A Primer on Quantile Regression for Epidemiologists

Aayush Khadka, Jilly Hebert, Anusha M. Vable

Department of Family and Community Medicine
University of California San Francisco

THE PREPRINT SERVER FOR HEALTH SCIENCES

Cold Spring Harbor Laboratory

Quantile regressions as a tool to evaluate how an exposure shifts and reshapes the outcome distribution: A primer for epidemiologists

Aayush Khadka, Jillian Hebert, M. Maria Glymour, Fei Jiang, Amanda Irish, Kate Duchowny, Anusha M. Vable

doi: https://doi.org/10.1101/2023.05.02.23289415

THE PREPRINT SERVER FOR HEALTH SCIENCES

Funding and acknowledgements

- Funding: National Institute on Aging (R01 AG069092, PI Vable).

- Many thanks to:
  - VabLab team
  - Maria Glymour’s research team
  - Dr. Catherine Duarte
  - ChatGPT

Overall learning aims

Theory	Practice
What is quantile regression and why do we care about it?	How do we run quantile regression analyses in R?
How does quantile regression differ from mean regression?	How can we present quantile regression coefficients?
How do estimators targeted at quantiles of the conditional vs. unconditional outcome distributions differ from one another?	How can we visualize how distributions are affected by an exposure using quantile regression results?

Overall learning aims

Theory	Practice
What is quantile regression and why do we care about it?	How do we run quantile regression analyses in R?
How does quantile regression differ from mean regression?	How can we present quantile regression coefficients?
How do estimators targeted at quantiles of the conditional vs. unconditional outcome distributions differ from one another?	How can we visualize how distributions are affected by an exposure using quantile regression results?

Overall learning aims

Theory	Practice
What is quantile regression and why do we care about it?	How do we run quantile regression analyses in R?
How does quantile regression differ from mean regression?	How can we present quantile regression coefficients?
How do estimators targeted at quantiles of the conditional vs. unconditional outcome distributions differ from one another?	How can we visualize how distributions are affected by an exposure using quantile regression results?

Previewing our practical example

- We will investigate the relationship between educational attainment and systolic blood pressure (SBP) in later-life
- Education has a strong, inverse relationship with average SBP levels
- There may be a non-linear relationship between later-life SBP and risk of coronary heart disease and stroke
  - Motivates the need to investigate if educational attainment affects the distribution of blood pressure

Previewing our practical example

- We will use data from the US Health and Retirement Study
  - Nationally representative, multi-cohort, biennially conducted longitudinal study of non-institutionalized US adults 50+ years
- Exposure: Self-reported total years of schooling (5-17 years)
- Outcome: First recorded SBP (2006-2018)
- Covariates:
  - Age (linear + quadratic)
  - Sex (Female/Male persons)
  - Race (Black/Latinx/White)
  - Southern birth (yes/no)
  - Mother's education (5-17)
  - Father's education (5-17)
  - SBP measurement year (2006-2018)

Teaching resources

- Slide deck: theory discussion
  - ★: Key point
- Handout: includes write-up on quantile regression + R code
- GitHub: includes the R Markdown file for the handout, additional code for the figures, slide deck, and other resources
  - Link: https://github.com/JillianHebert/A-Primer-on-Quantile-Regression-for-Epidemiologists

Terminology we use interchangeably

Marginal and unconditional

Terminology we use interchangeably

Marginal and unconditional

Means and expectation

Terminology we use interchangeably

Marginal and unconditional

Means and expectation

Scale, spread, and variance

Terminology we use interchangeably

Marginal and unconditional

Means and expectation

Scale, spread, and variance

Epsilon and error

Terminology we use interchangeably

Marginal and unconditional

Means and expectation

Scale, spread, and variance

Epsilon and error

Tangerine and orange

Terminology we use interchangeably

Marginal and unconditional

Means and expectation

Scale, spread, and variance

Epsilon and error

Tangerine and orange

Teal and green

Terminology we use interchangeably

Marginal and unconditional

Means and expectation

Scale, spread, and variance

Epsilon and error

Tangerine and orange

Teal and green

Rose and red

Abbreviations

Abbreviation	Full form
CDF	Cumulative Distribution Function
CEF	Conditional Expectation Function
CQF	Conditional Quantile Function
CQR	Conditional Quantile Regression
DGP	Data Generating Process
IF	Influence function
OLS	Ordinary Least Squares
RIF	Recentered Influence Function
SBP	Systolic Blood Pressure
UQR	Unconditional Quantile Regression

Outline of the workshop

1. Importance of focusing on the entire outcome distribution (~12 mins)
2. Means and quantiles (~30 mins) 10-minute break
3. Linear regression + R code (~30 mins) 30-minute food break
4. Conditional quantile regression + R code (~45 mins) 10-minute break
5. Unconditional quantile regression + R code (~45 mins)
6. Compare and conclude (~15 mins)

Key takeaways from this workshop

1. Investigating how an exposure affects the entire outcome distribution, in particular the tails, is substantively important

2. Quantile regressions allow us to quantify the relationship of an exposure with the outcome distribution

3. Need to determine if we are interested in quantiles of the conditional or unconditional outcome distribution in advance

4. Separate estimators need to be used for quantiles of the conditional versus unconditional outcome distribution

A matter of life and distributions

i.e., a case for why we should focus on the entire outcome distribution

Learning aims

1. Why it’s important to think about the entire outcome distribution

2. When can mean models help us learn about how an exposure affects the entire outcome distribution?

“Geoffrey Rose’s big idea”*

Sick individuals and sick populations

Geoffrey Rose

Rose G (Department of Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK). Sick individuals and sick populations. *International Journal of Epidemiology* 1985;**14**:32–38.

Aetiology confronts two distinct issues: the determinants of individual cases, and the determinants of incidence rate. If exposure to a necessary agent is homogeneous within a population, then case/control and cohort methods will fail to detect it: they will only identify markers of susceptibility. The corresponding strategies in control are the ‘high-risk’ approach, which seeks to protect susceptible individuals, and the population approach, which seeks to control the causes of incidence. The two approaches are not usually in competition, but the prior concern should always be to discover and control the causes of incidence.

*The title of this slide comes from a Commentary written by A Hofman and JP Vandenbroucke (1992). Link

Determinants of cases vs. determinants of incidence

	Determinants of cases	Determinants of incidence
Question of interest	Why did this patient get this disease at this time?	Why do some populations, but not others, have high rates of disease?

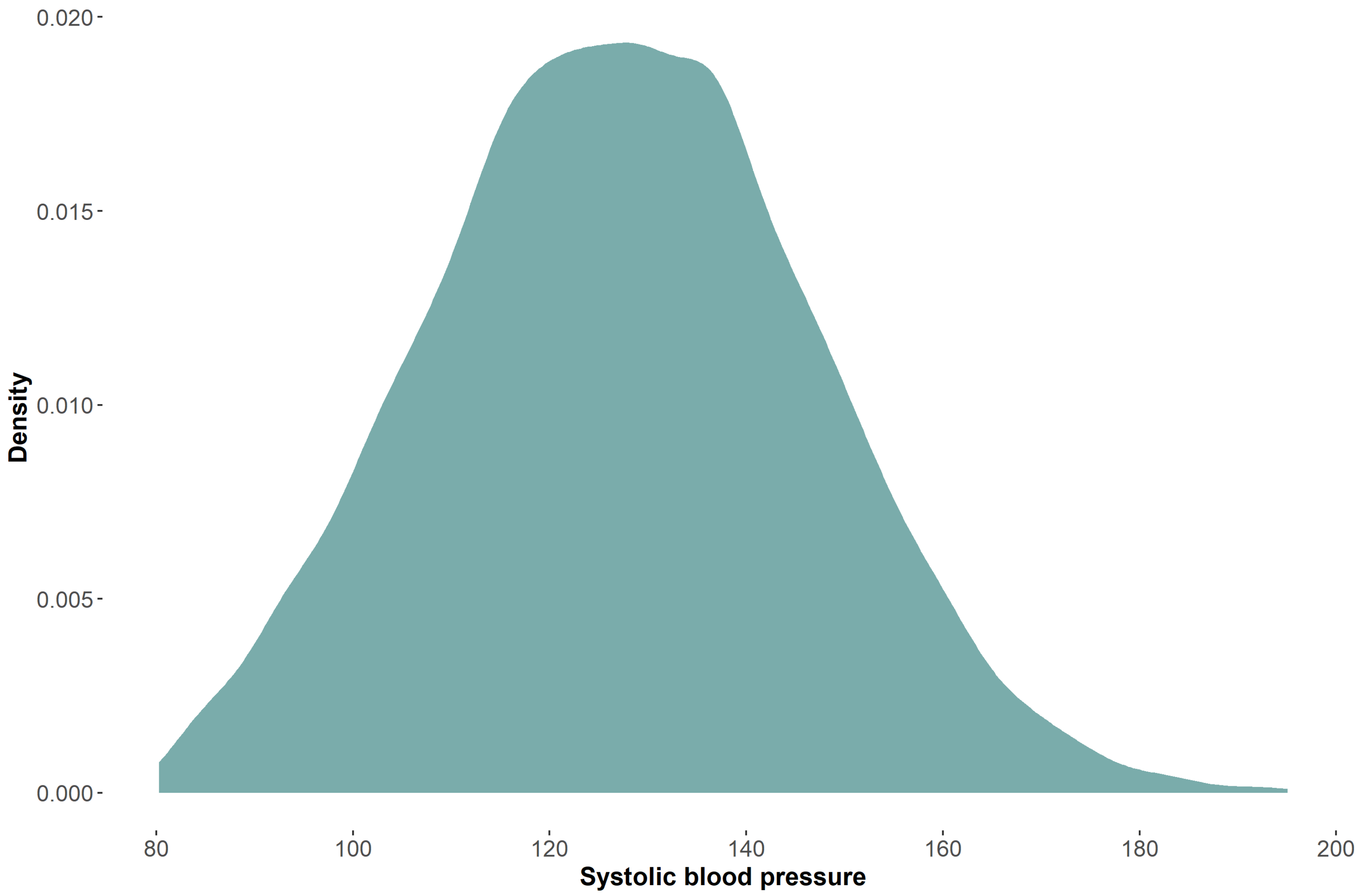
Determinants of cases vs. determinants of incidence

	Determinants of cases	Determinants of incidence
Question of interest	Why did this patient get this disease at this time?	Why do some populations, but not others, have high rates of disease?
Approach to answering question of interest	Search for individual-level risk factors	Search for population-level determinants of disease distribution

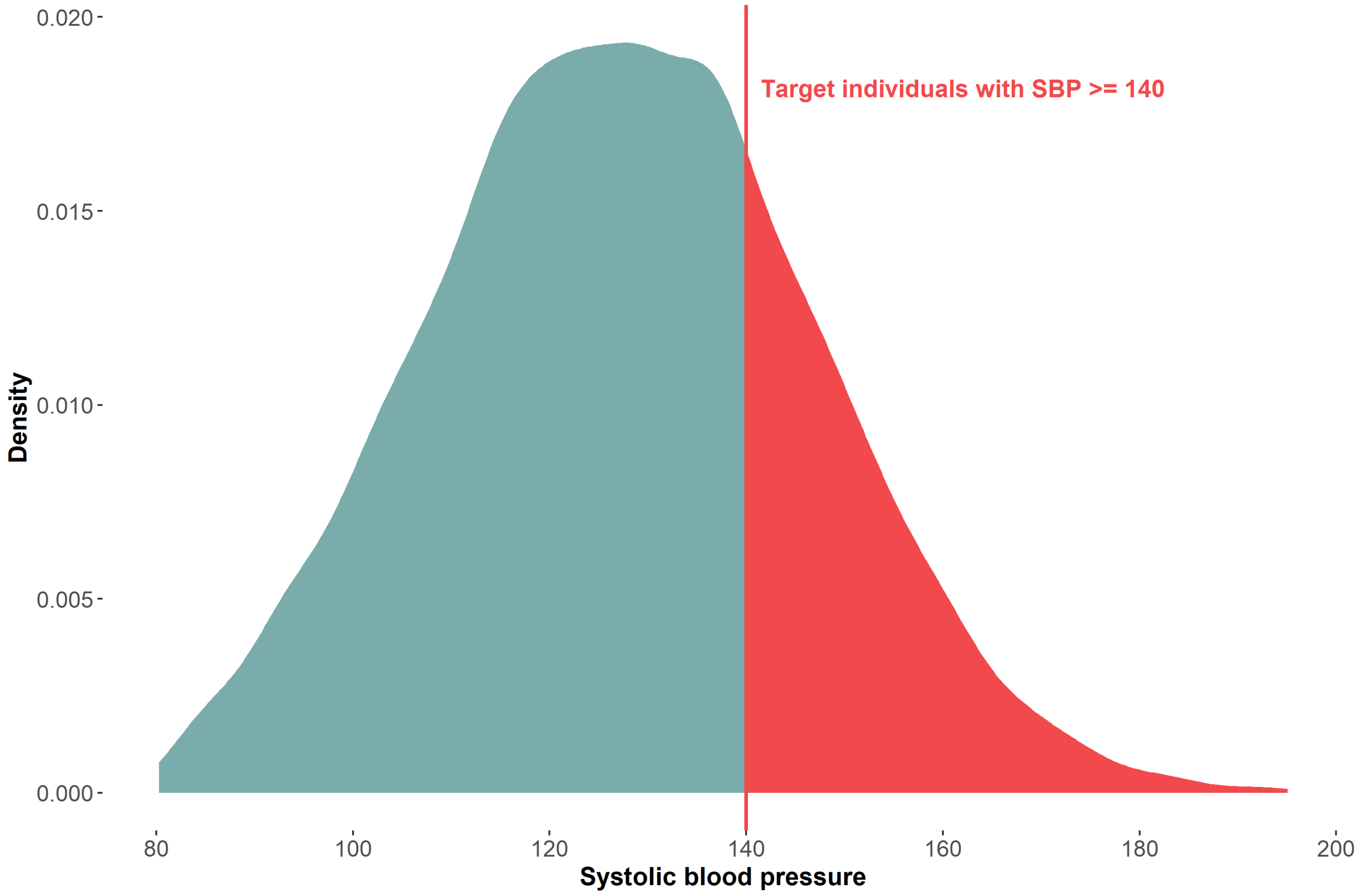
Determinants of cases vs. determinants of incidence

	Determinants of cases	Determinants of incidence
Question of interest	Why did this patient get this disease at this time?	Why do some populations, but not others, have high rates of disease?
Approach to answering question of interest	Search for individual-level risk factors	Search for population-level determinants of disease distribution
Implied approach for reducing disease burden	Target individuals at high risk of disease	Shift the distribution of the disease determinant in the population

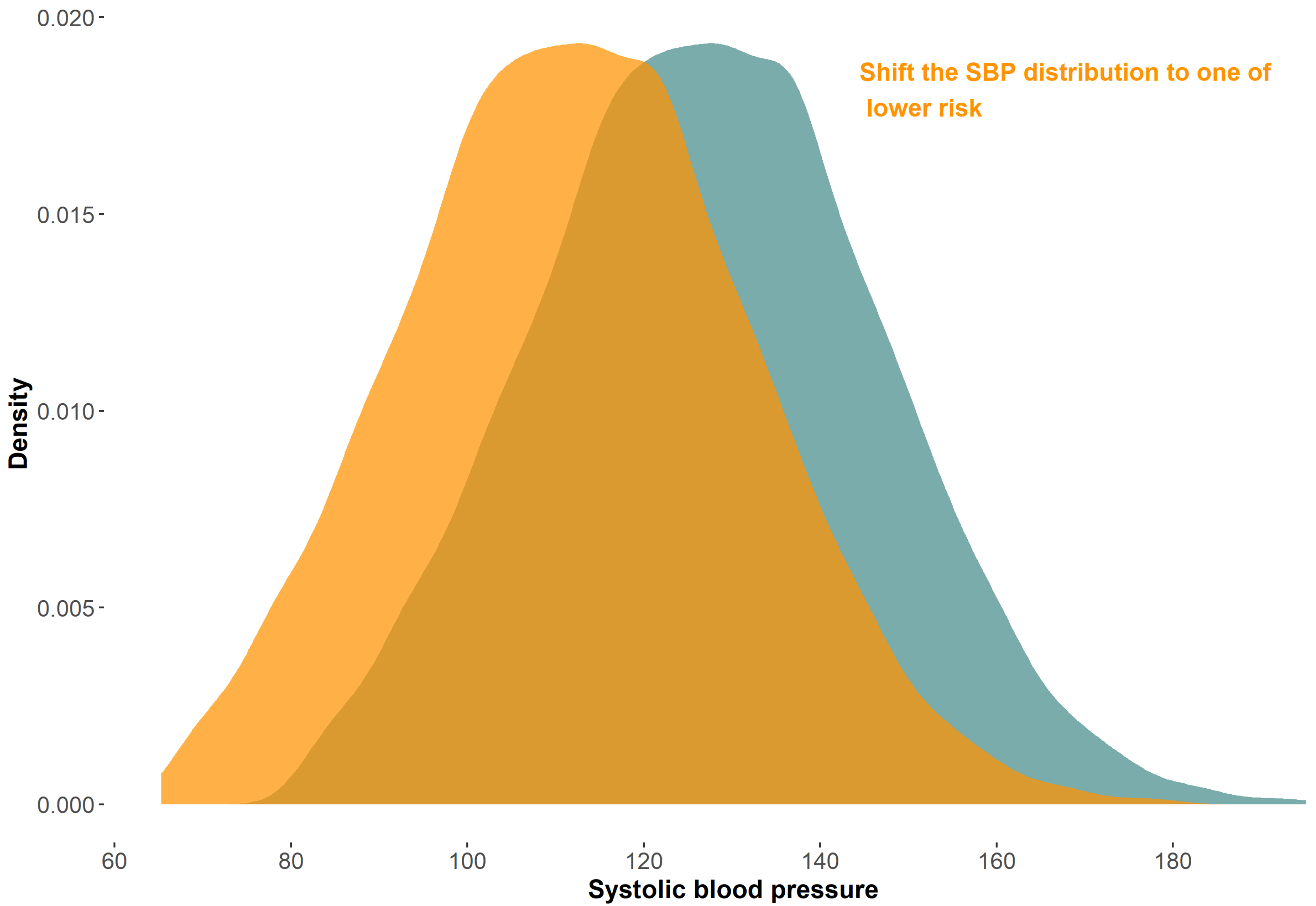
Illustrating Rose's big idea



High risk approach for prevention



Population approach for prevention

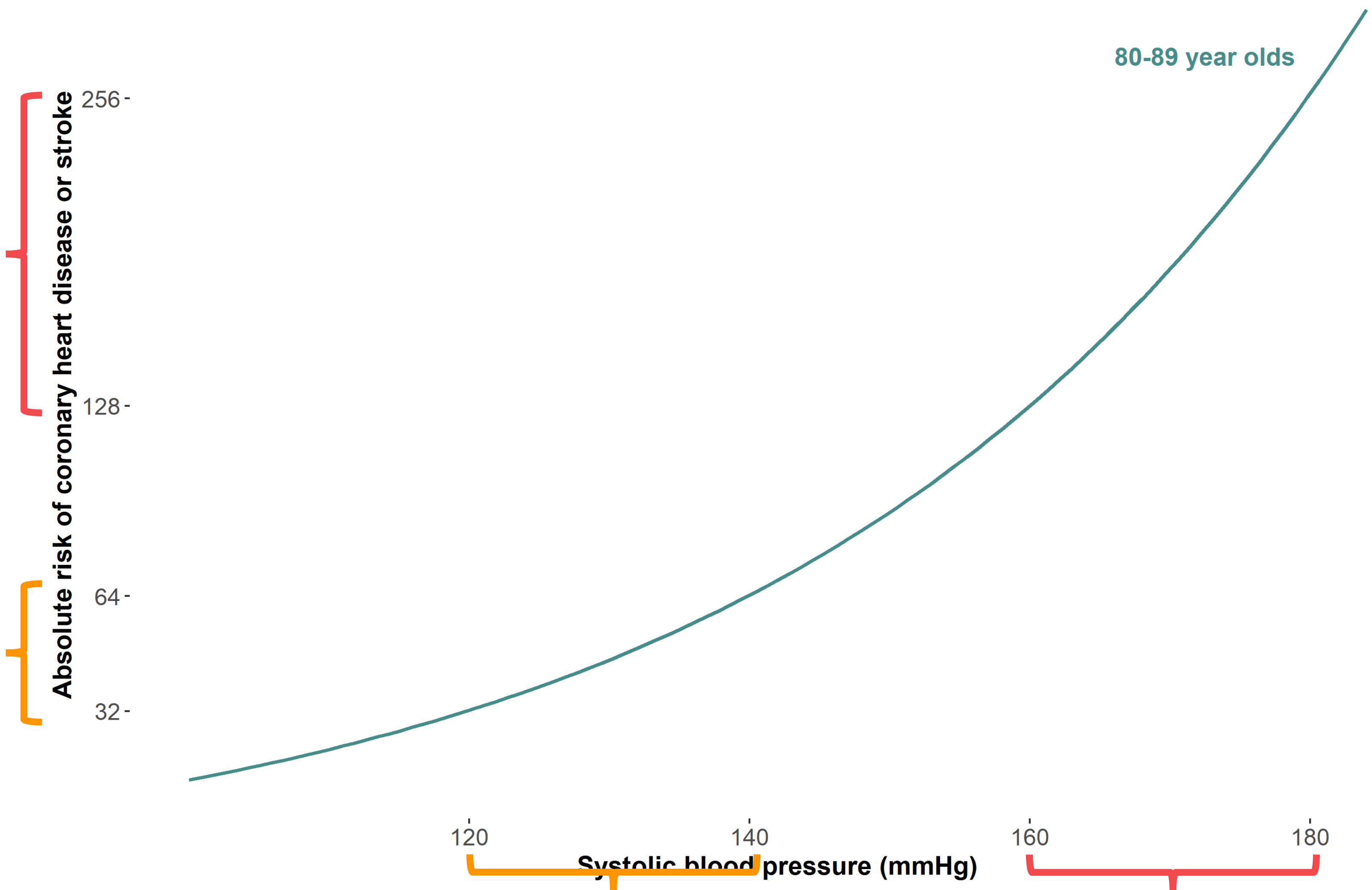


Rose's framework privileges the mean of a risk factor's distribution

The study of individual susceptibility is important, but the fundamental questions are, "What determines the population's mean blood pressure, cholesterol, body weight, and alcohol intake? And how might that mean be changed?"

- Geoffrey Rose (1991). Ancel Keys lecture

The tails matter too: case of blood pressure



This figure is a recreation of a figure from Fuchs et. al. (2020 with simulated data.

Empirical epidemiology focuses on modeling the outcome mean

- Epidemiologists are not exposed to regression models for quantities other than the mean value of an outcome

- For a binary outcome, the mean describes the distribution

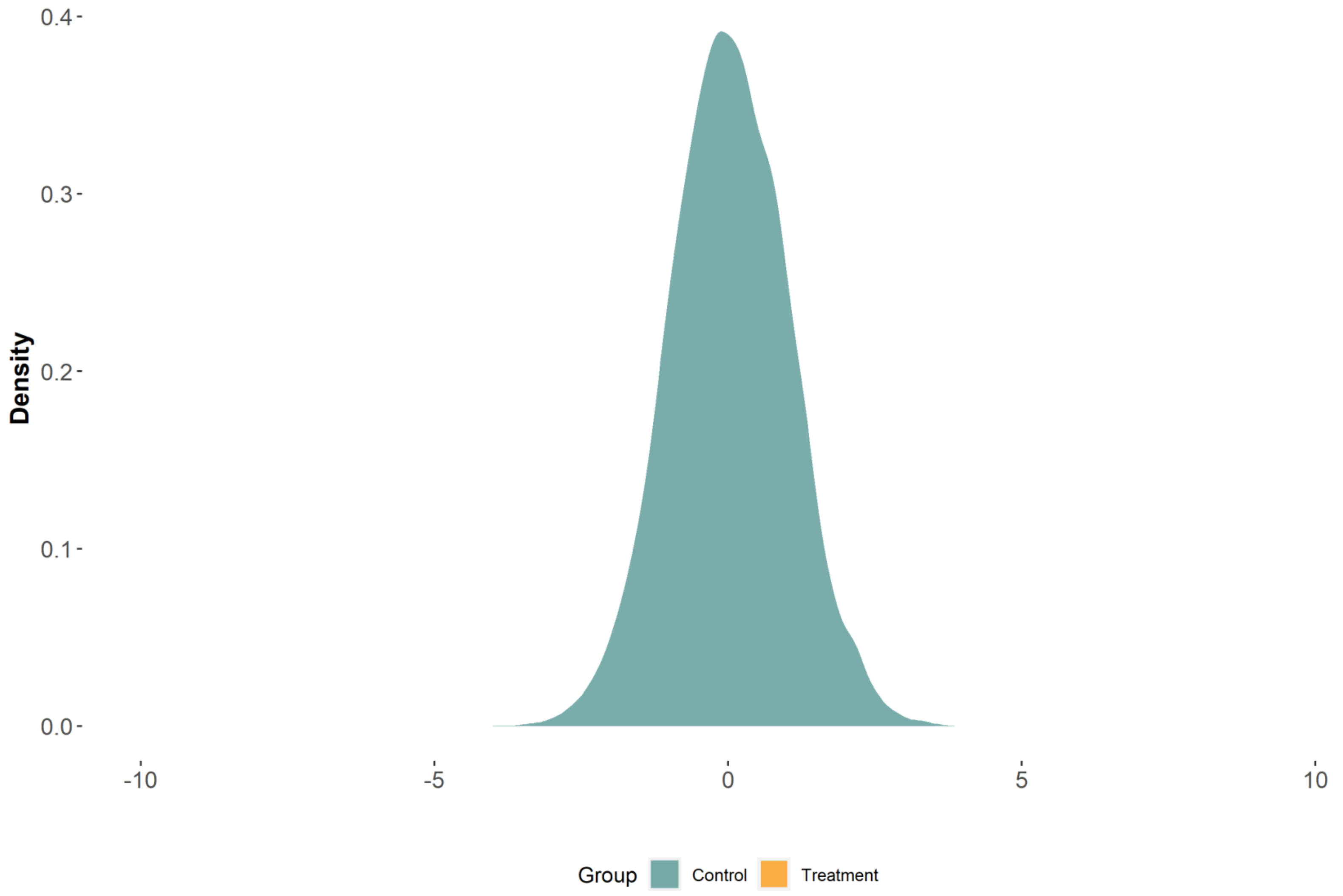
- Output from models not focused on the outcome mean may be trickier to interpret

Can mean models quantify how an exposure affects different parts of the outcome distribution?

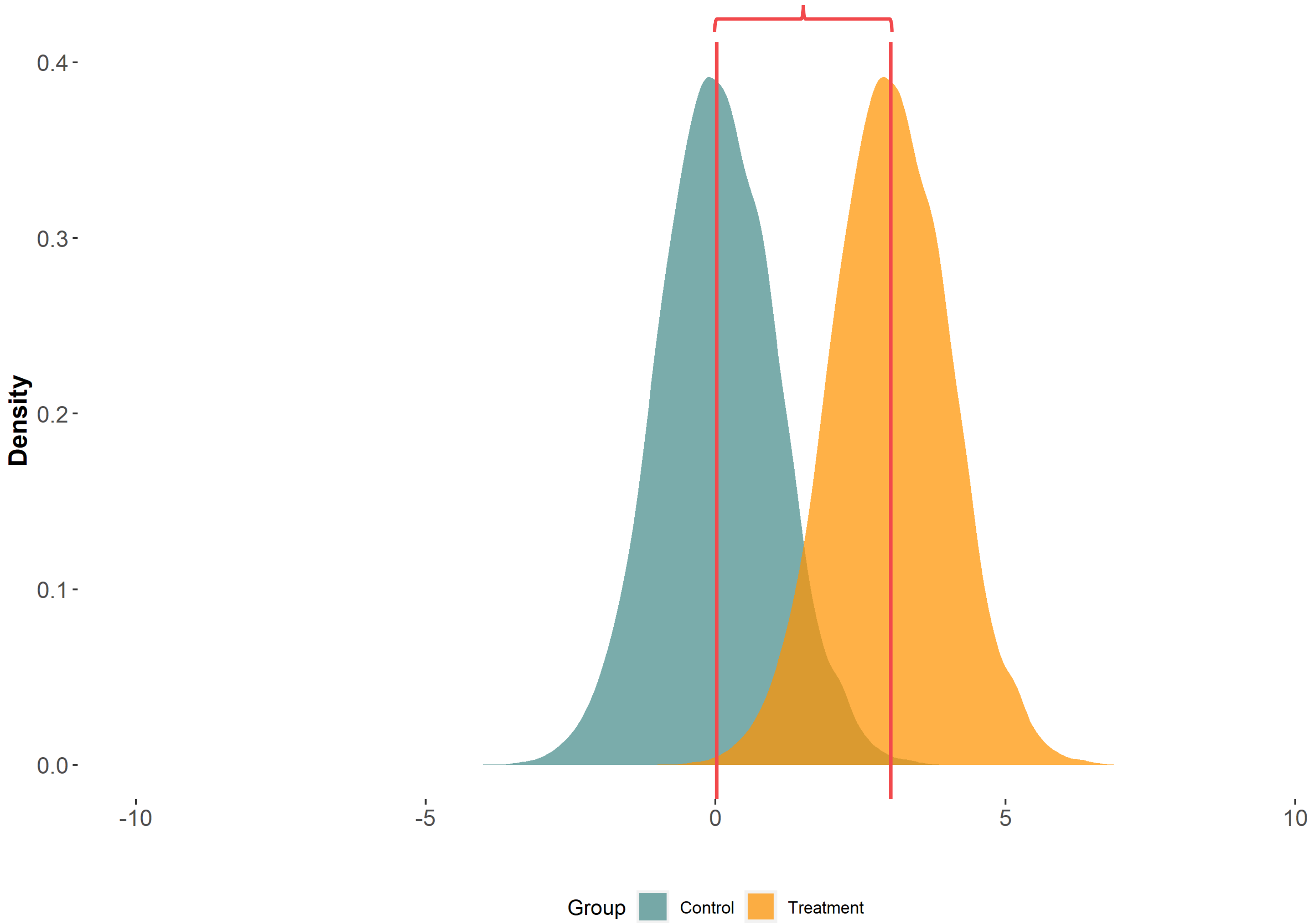
Means may tell tall tales about tails

(repeat 5 times fast)

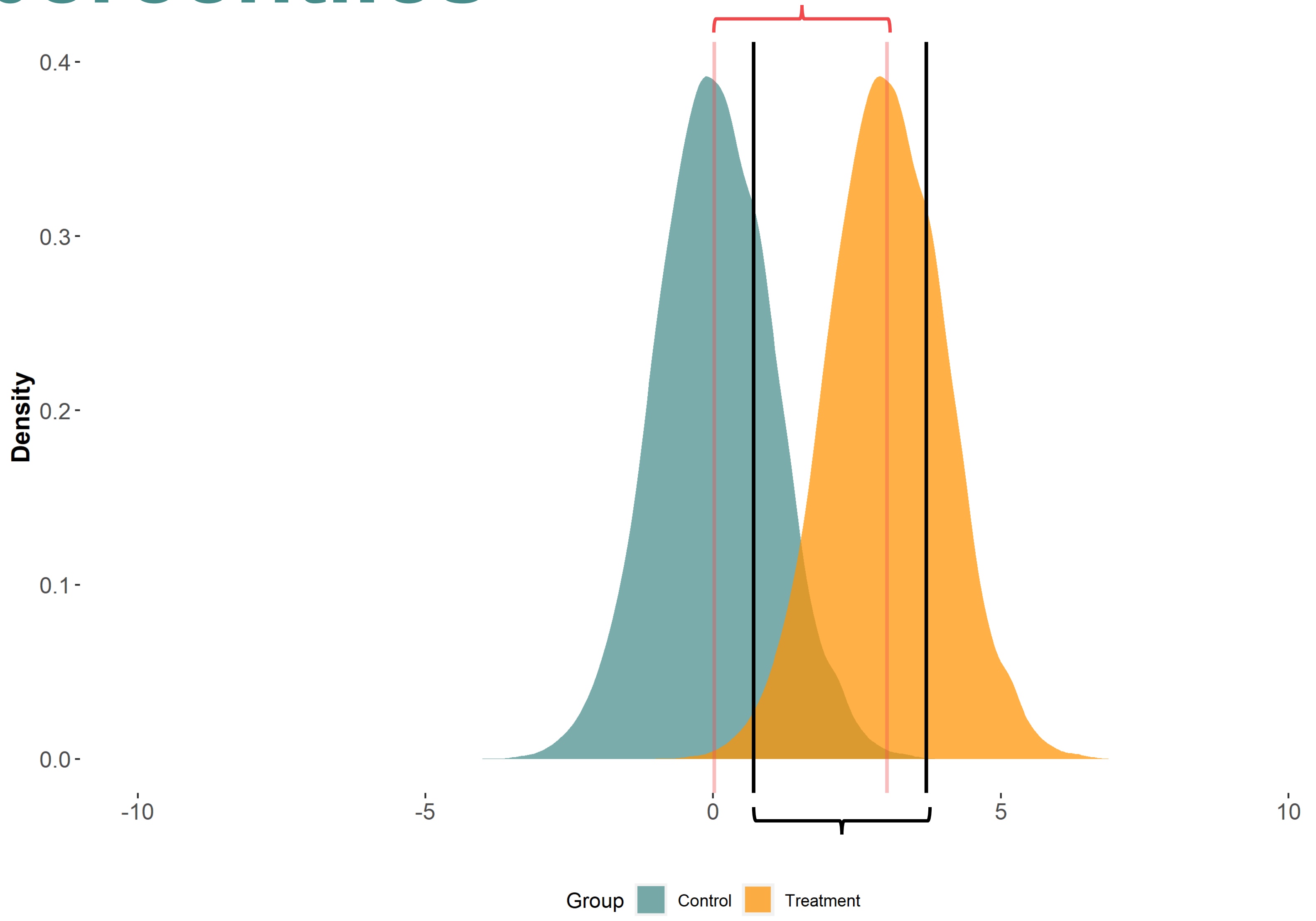
Imagine that a treatment shifts the outcome distribution uniformly



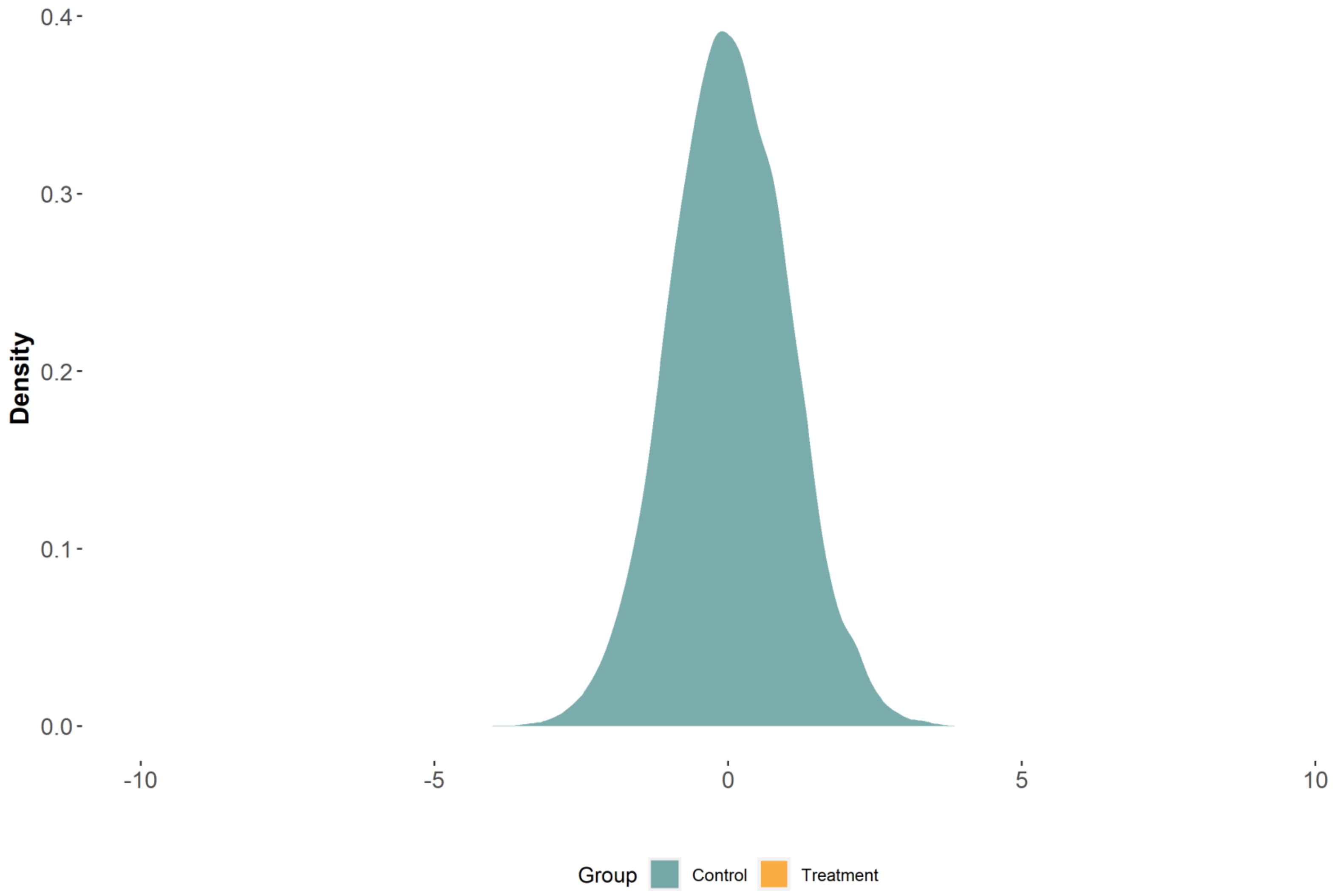
Difference in means



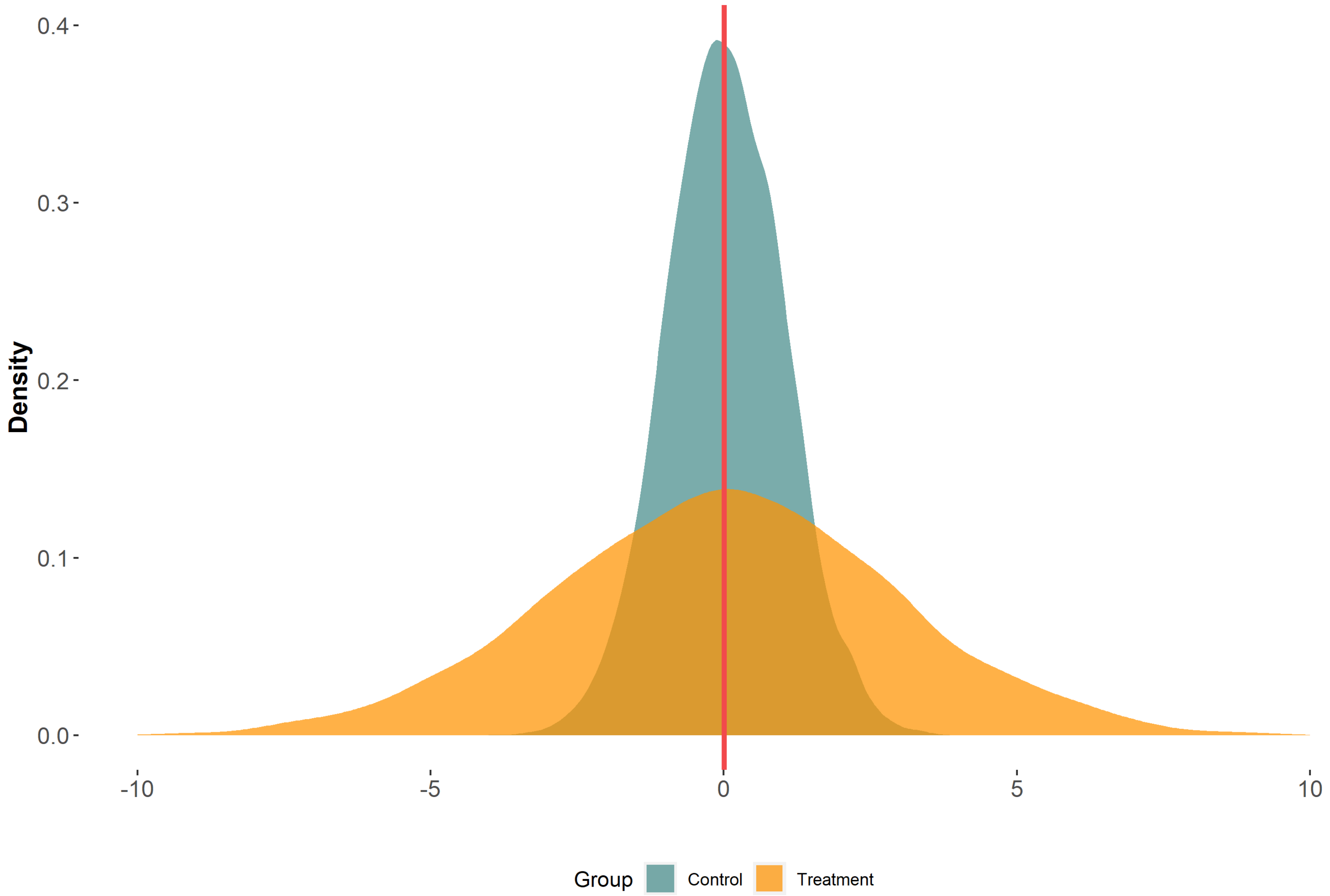
Difference in means = difference in 75^{th} percentiles



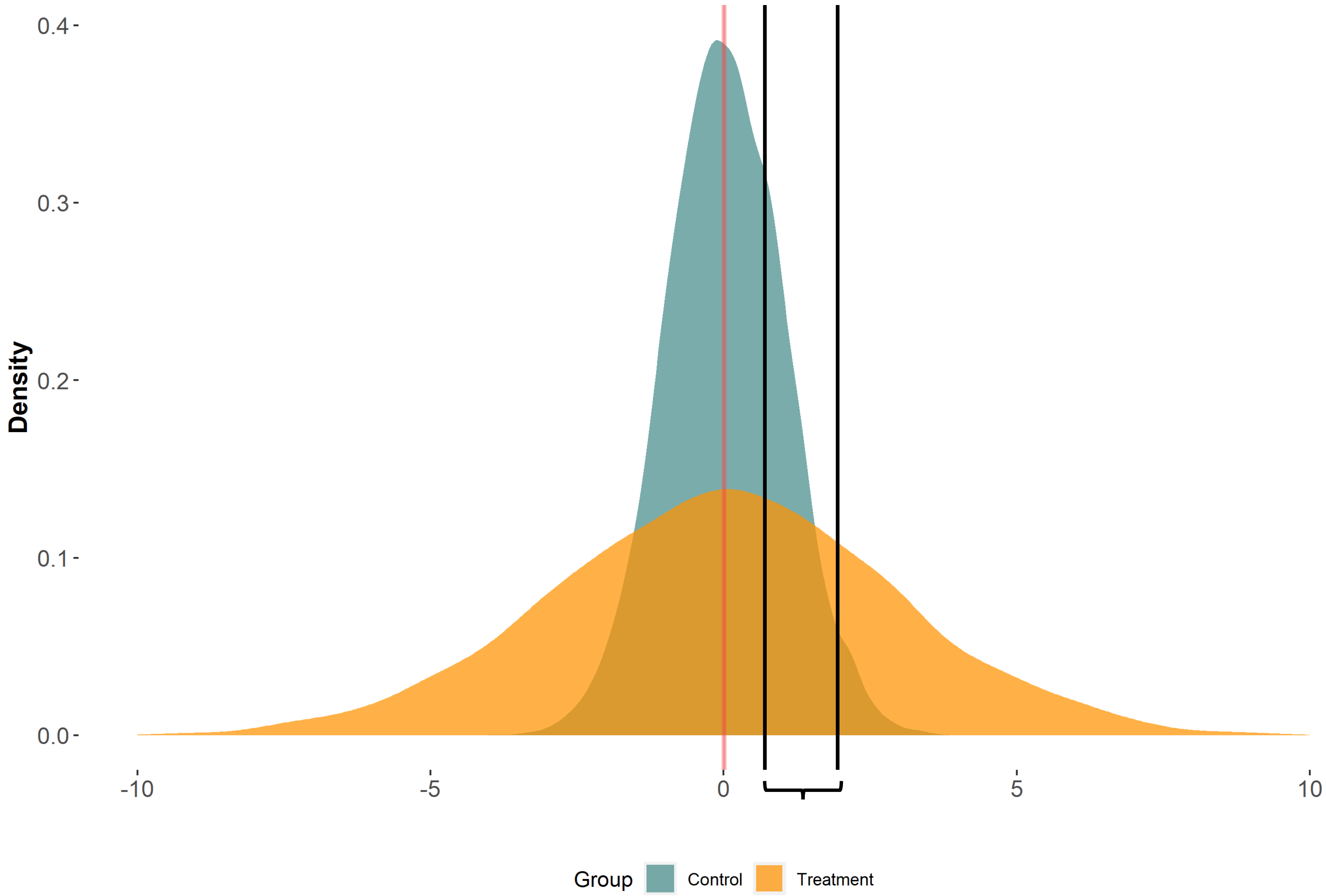
Imagine that a treatment affects the outcome distribution's scale only



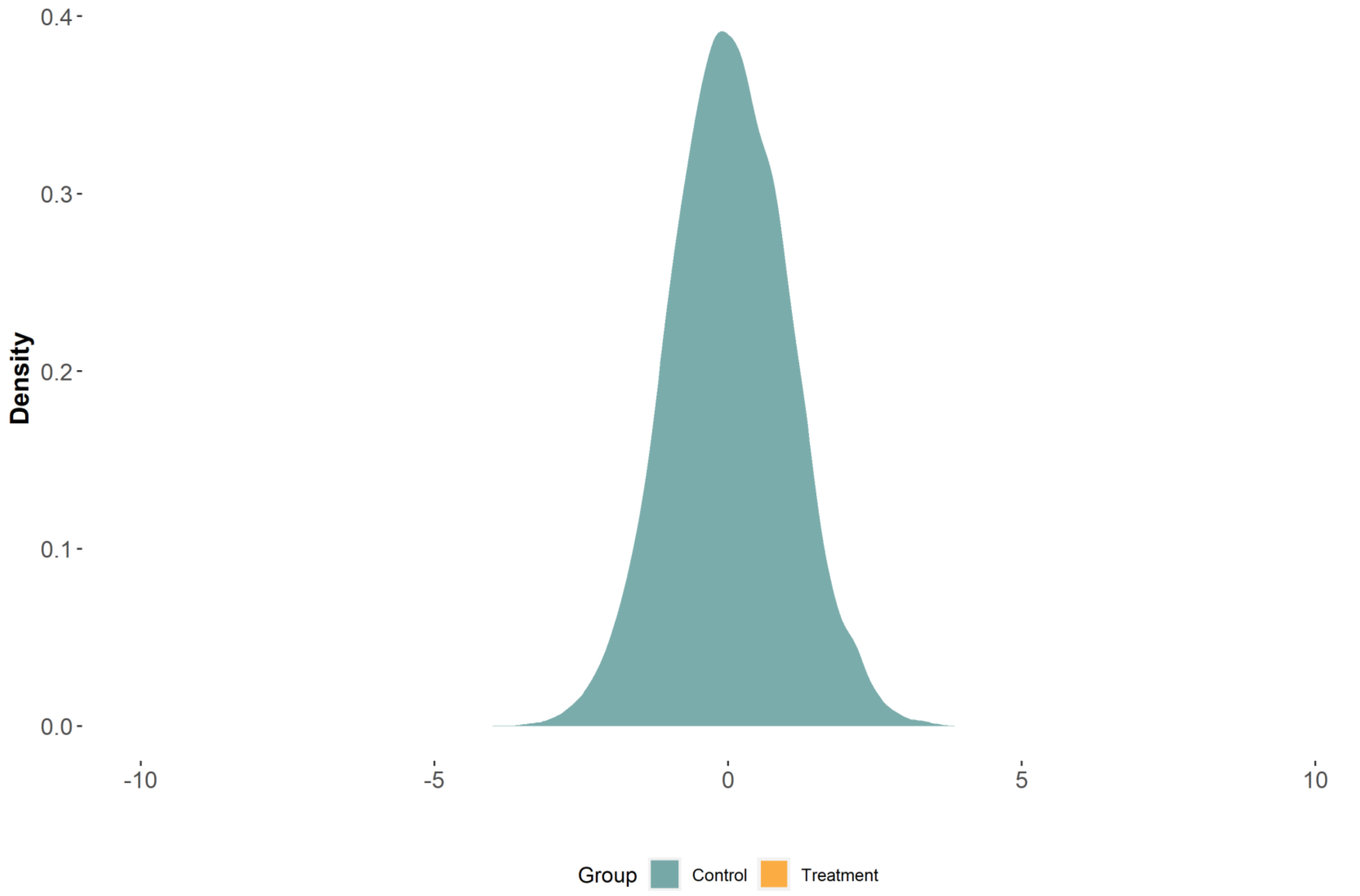
Difference in means



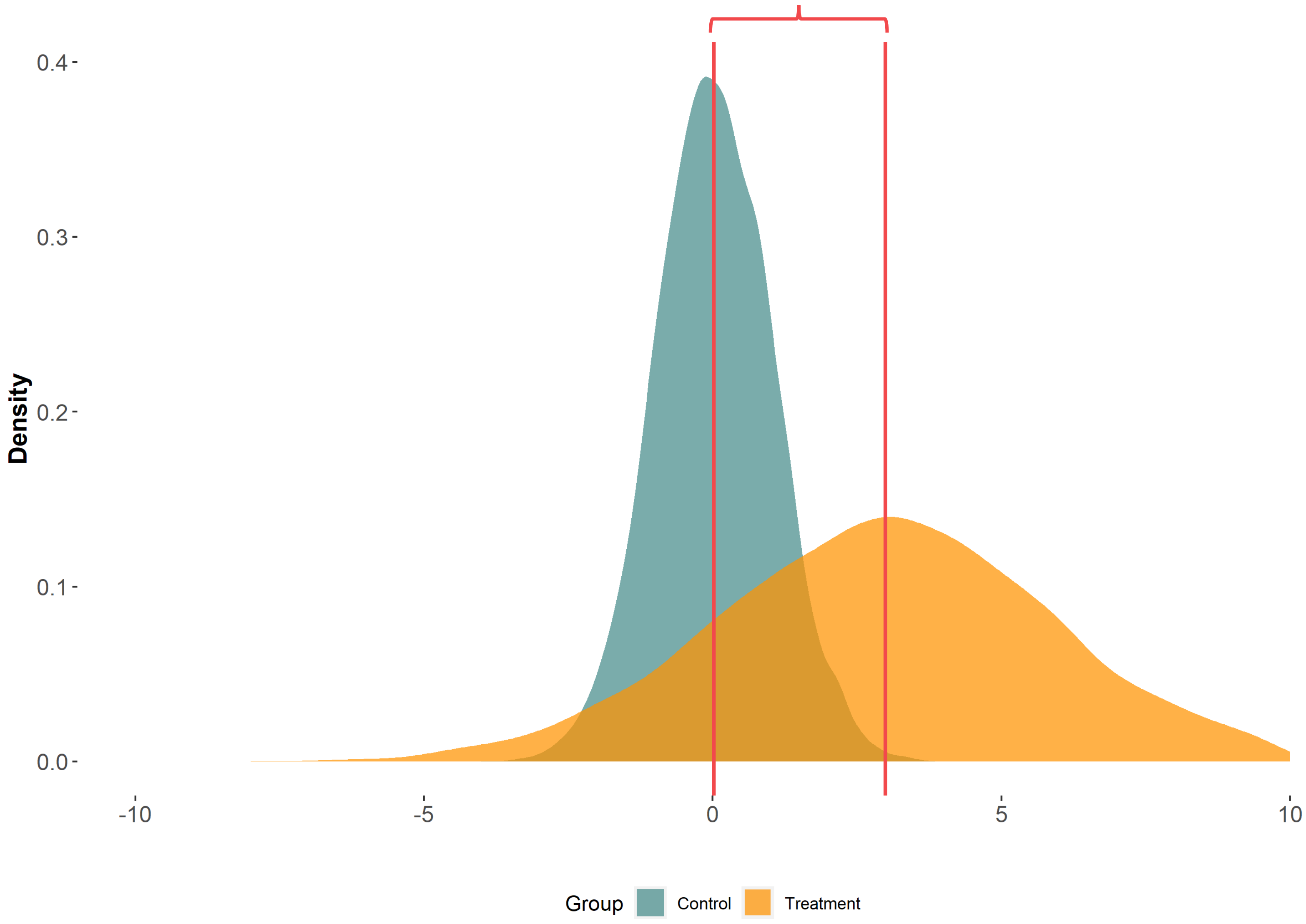
Difference in means ≠ difference in 75^{th} percentiles



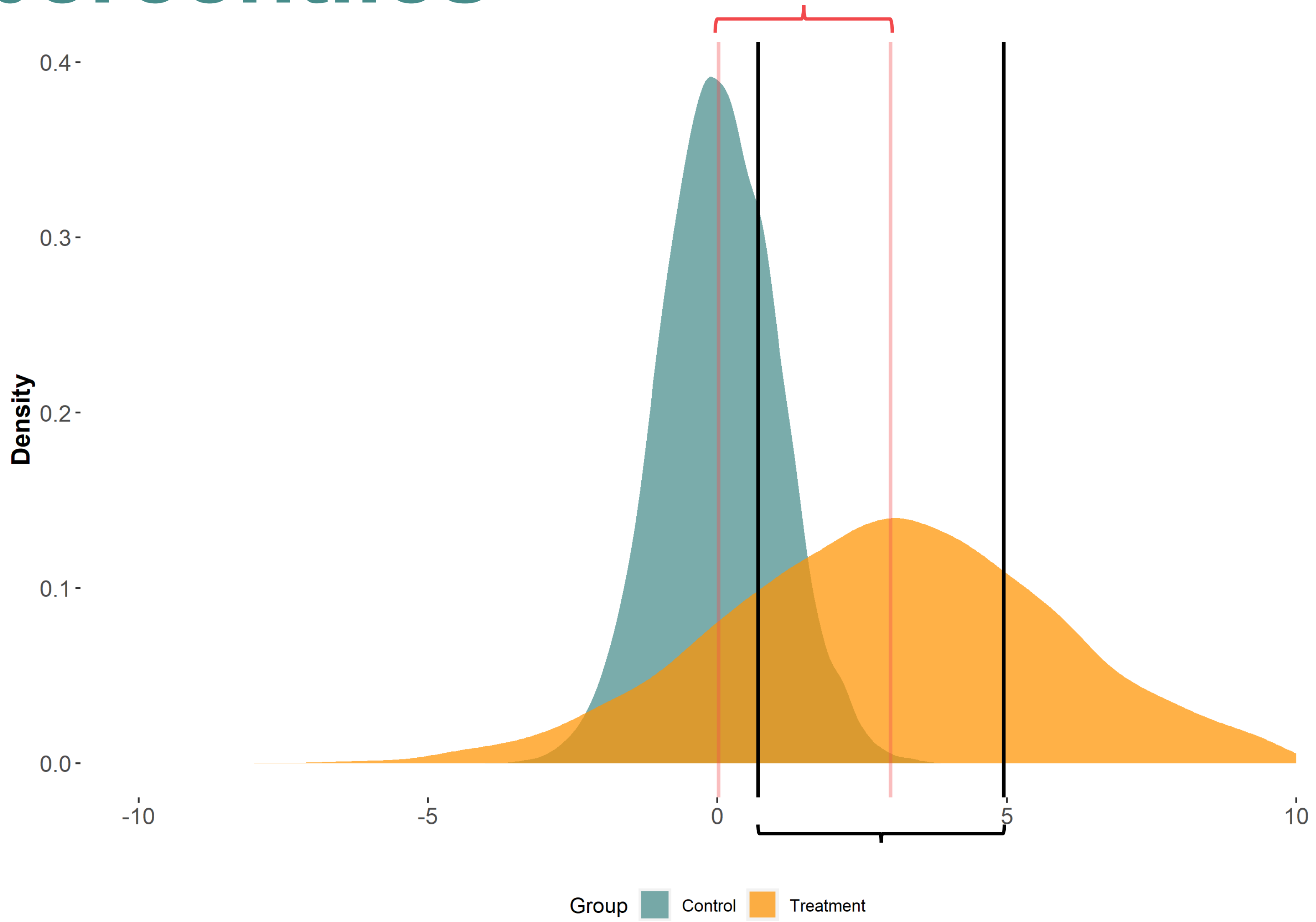
Imagine that a treatment affects both the location and scale



Difference in means



Difference in means ≠ difference in 75th percentiles



Mean models cannot quantify distributional effects when a treatment affects the scale of the outcome distribution

Key takeaways

- Investigating how an exposure affects different parts of the outcome distribution, esp. the tails, may be important
  - Most vulnerable individuals may be in the tails of the distribution
- Empirical epidemiology focuses primarily on modeling the outcome mean
- Mean models are limited in their ability to quantify effects across the outcome distribution

A brewing battle between means and quantiles

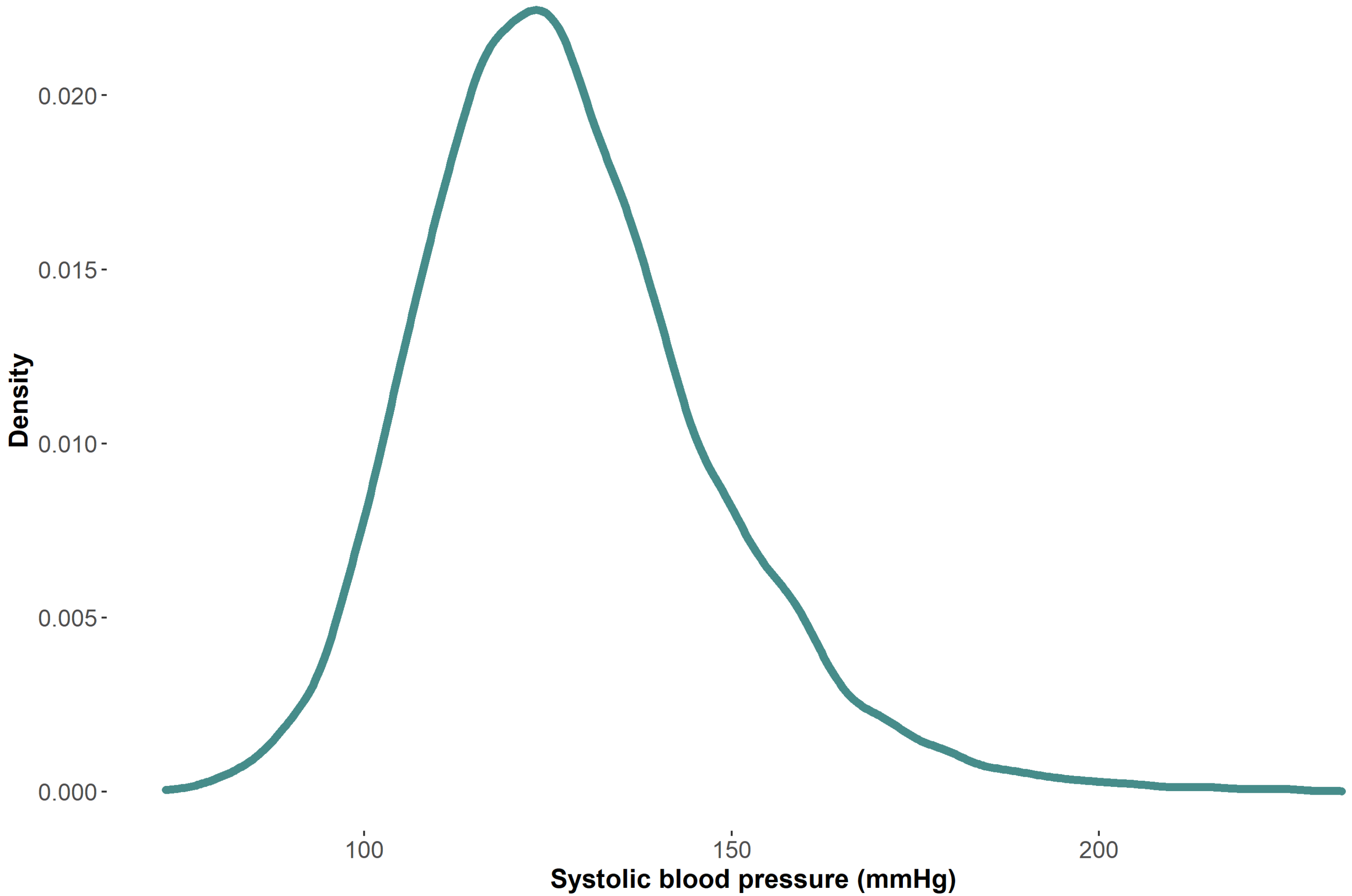
i.e., a review of means and quantiles

Learning aims

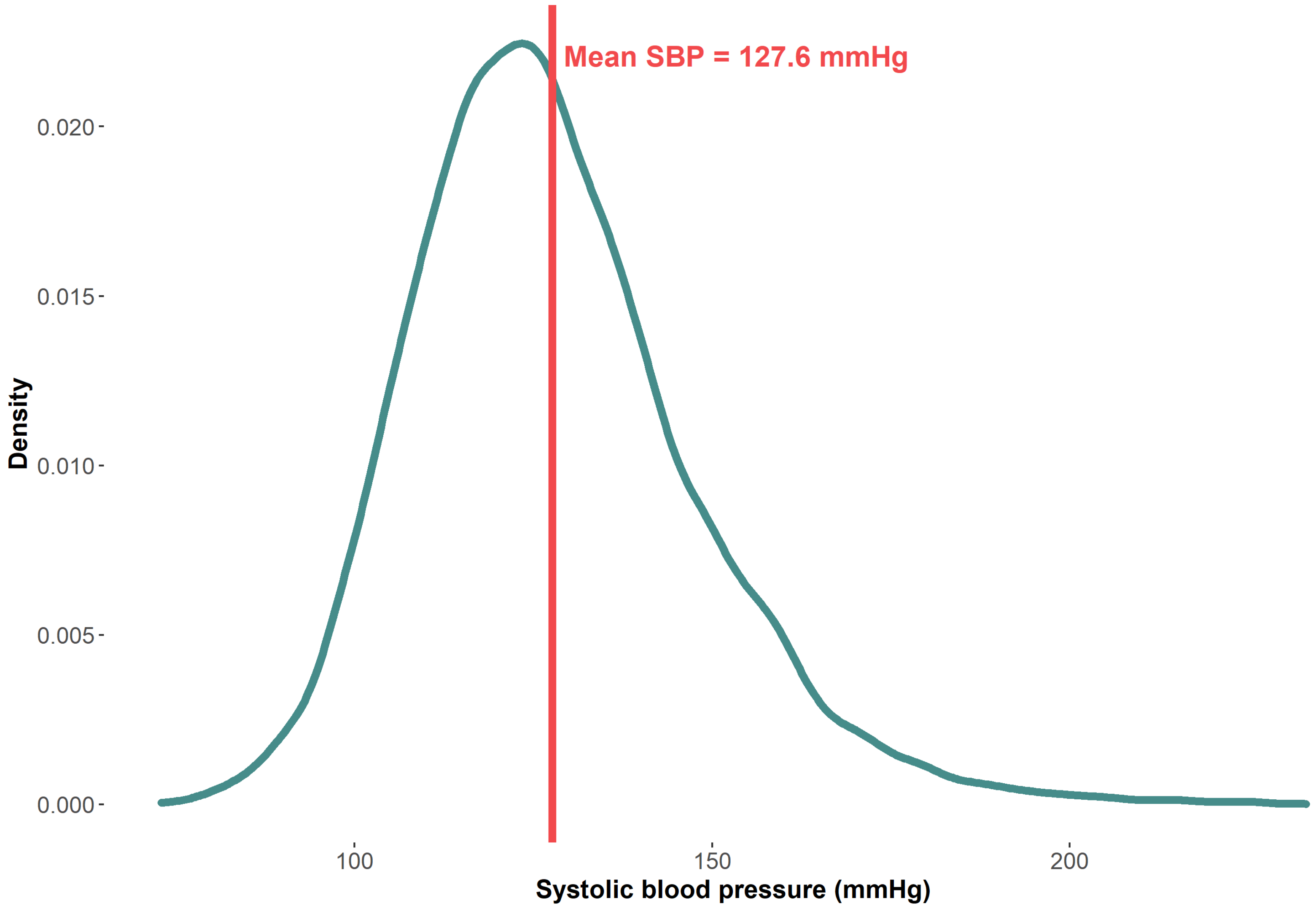
1. Mean:
   a. Unconditional and conditional means (concept + estimation)
   b. Law of Iterated Expectations

2. Quantiles
   a. Unconditional and conditional quantiles (concept + estimation)
   b. Law of Iterated Quantiles?

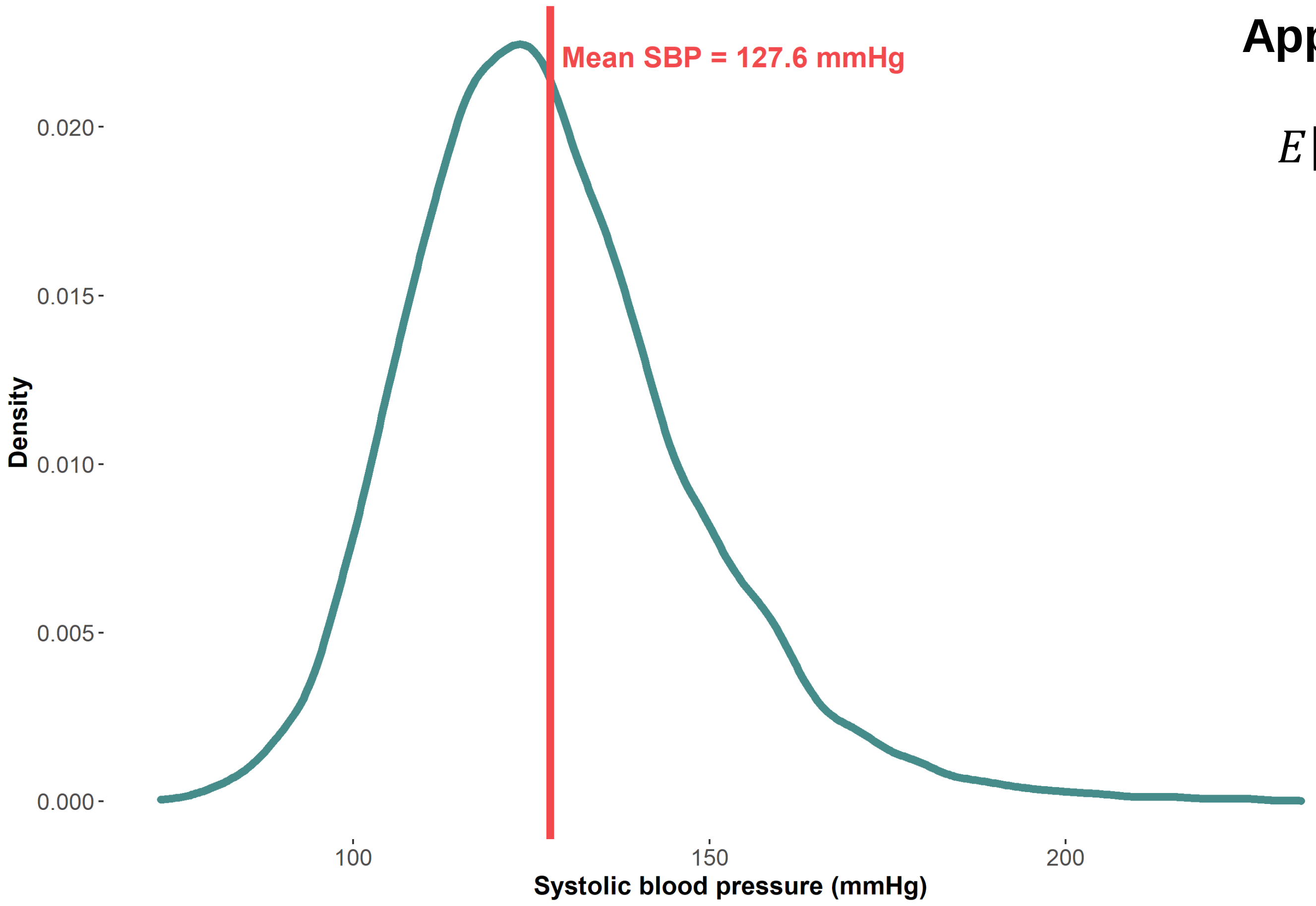
Empirical unconditional distribution of SBP in our data



(Arithmetic) Mean of the empirical unconditional SBP distribution



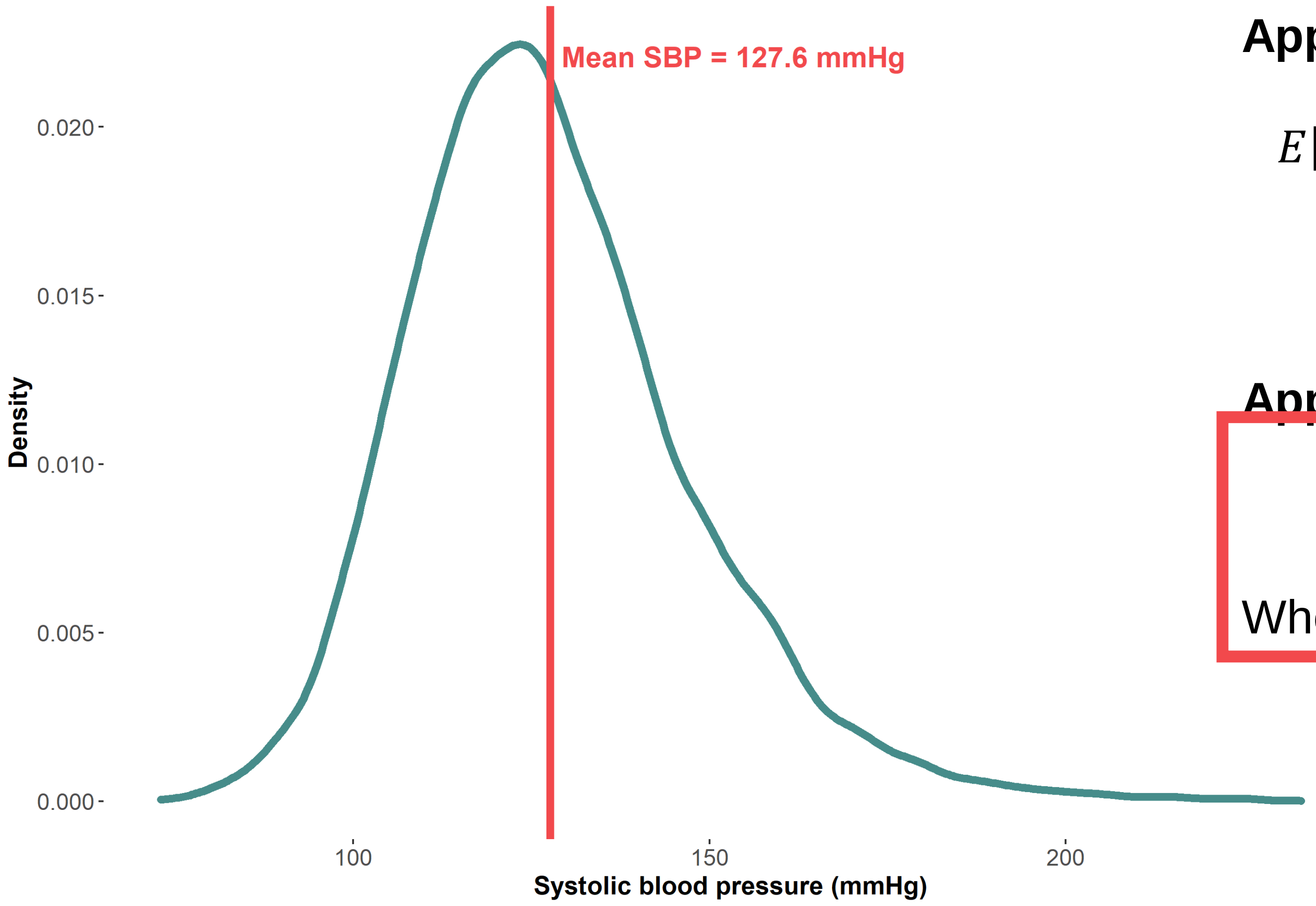
How can we estimate the mean of the unconditional SBP distribution?



Approach 1:

$$E[SBP] = \boxed{\sum_{i=1}^{N} SBP_i * \Pr(SBP_i)} = \boxed{\frac{1}{N}\sum_{i=1}^{N} SBP_i}$$

How can we estimate the mean of the unconditional SBP distribution?



Approach 1:

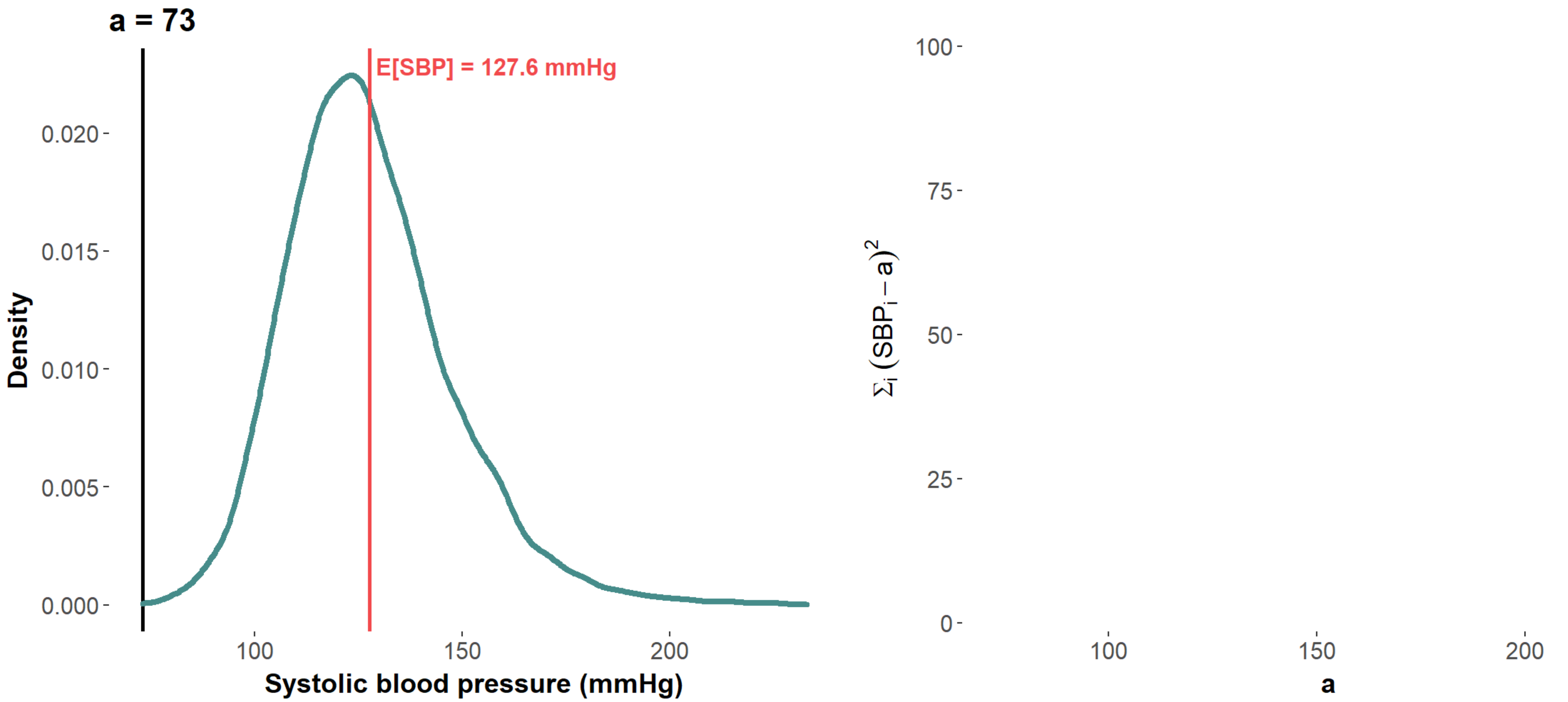
$$E[SBP] = \sum_{i=1}^{N} SBP_i * \Pr(SBP_i) = \frac{1}{N}\sum_{i=1}^{N} SBP_i$$

Approach 2:

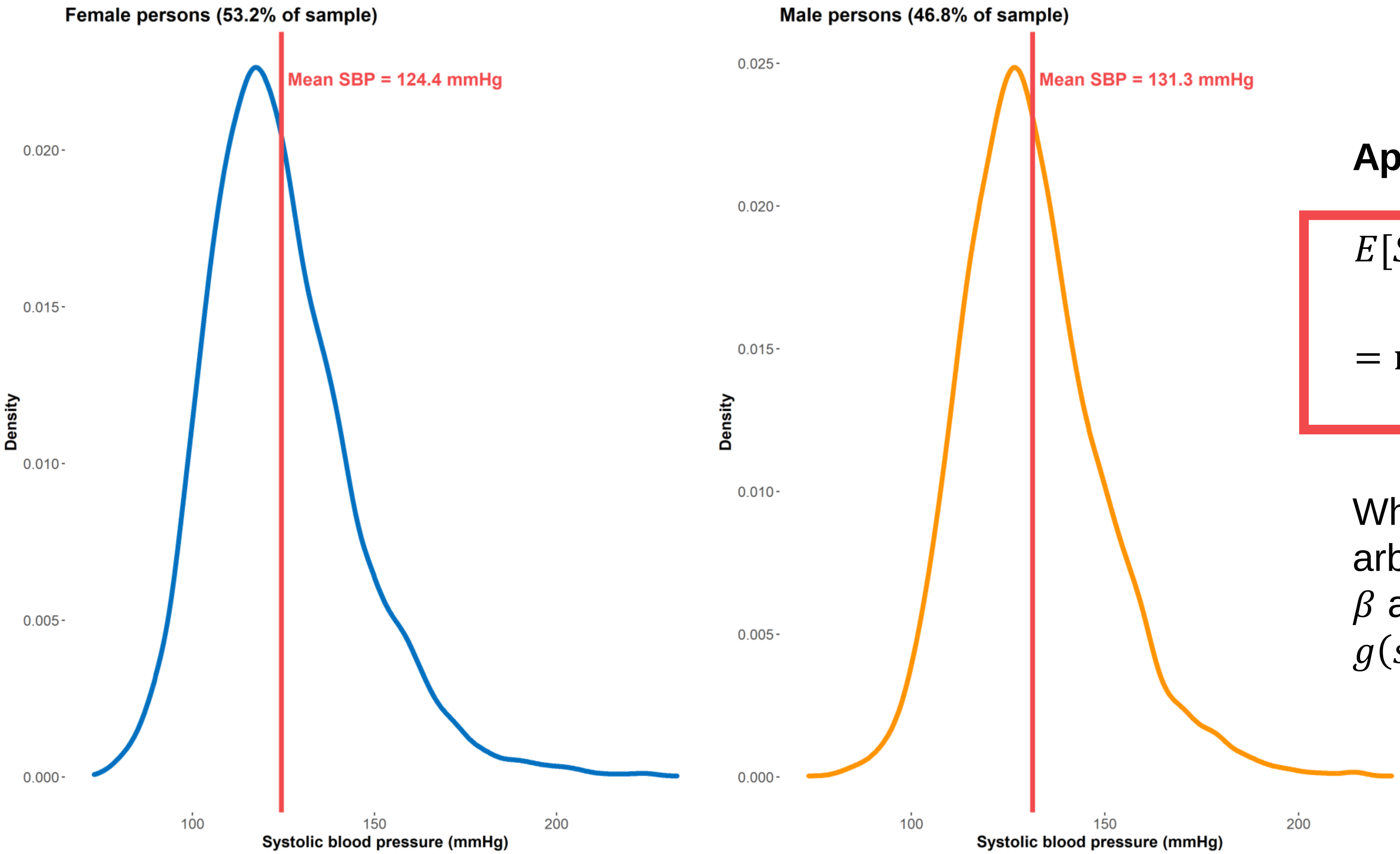
$$E[SBP] = \min_{a} \sum_{i=1}^{N} (SBP_i - a)^2$$

Where a is any value in the range of SBP

Visualizing $E[SBP] = \min_a \sum_{i=1}^{N} (SBP_i - a)^2$



Mean SBP conditional on self-identified sex

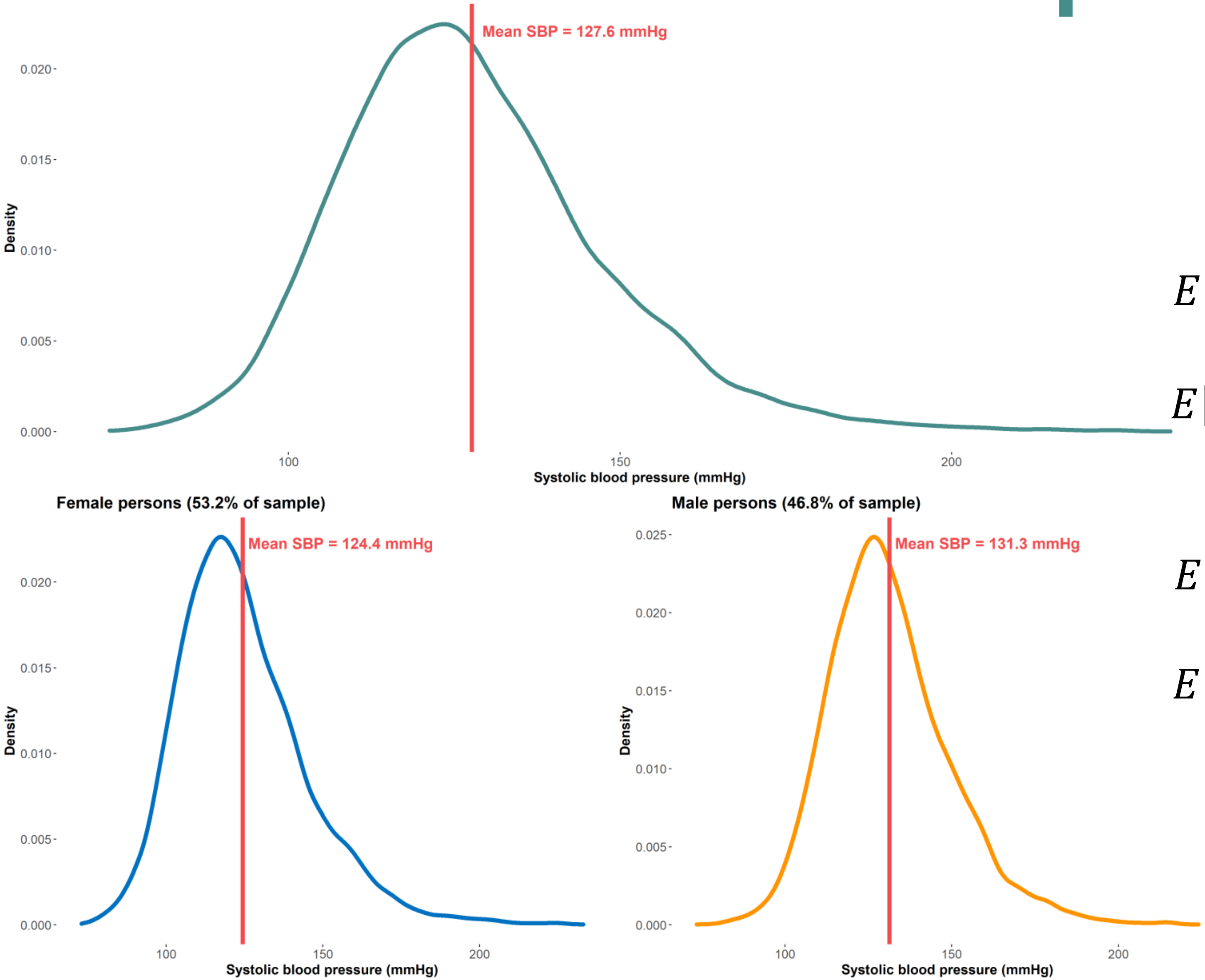


Approach:

$$\boxed{E[SBP|sex] = \min_{\beta}\sum_{i=1}^{N}\left(SBP_i - g(sex_i,\beta)\right)^2}$$

Where $g(sex_i, \beta)$ is an arbitrary function of sex with β as its parameters (e.g., $g(sex_i, \beta) = \beta_0 + \beta_1 sex_i$)

The Law of Iterated Expectations



$$E[SBP] = E_{sex}\big[E[SBP|sex]\big]$$

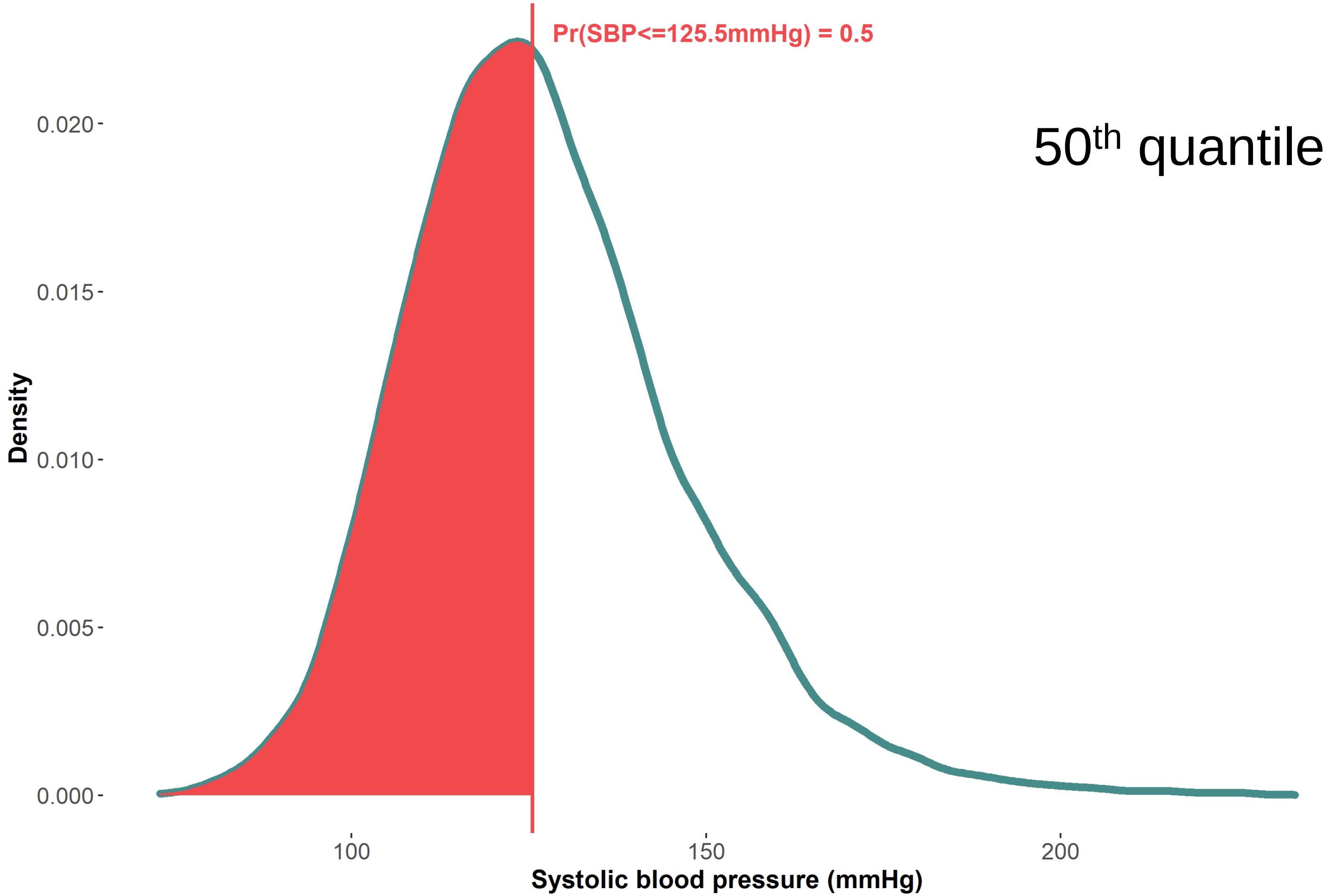
$$E[SBP] = \Pr(f) * E[SBP|sex = f] + \Pr(m) * E[SBP|sex = m]$$

$$E[SBP] = 0.532 * 124.4 + 0.468 * 131.3$$

$$E[SBP] = 127.6\ mmHg$$

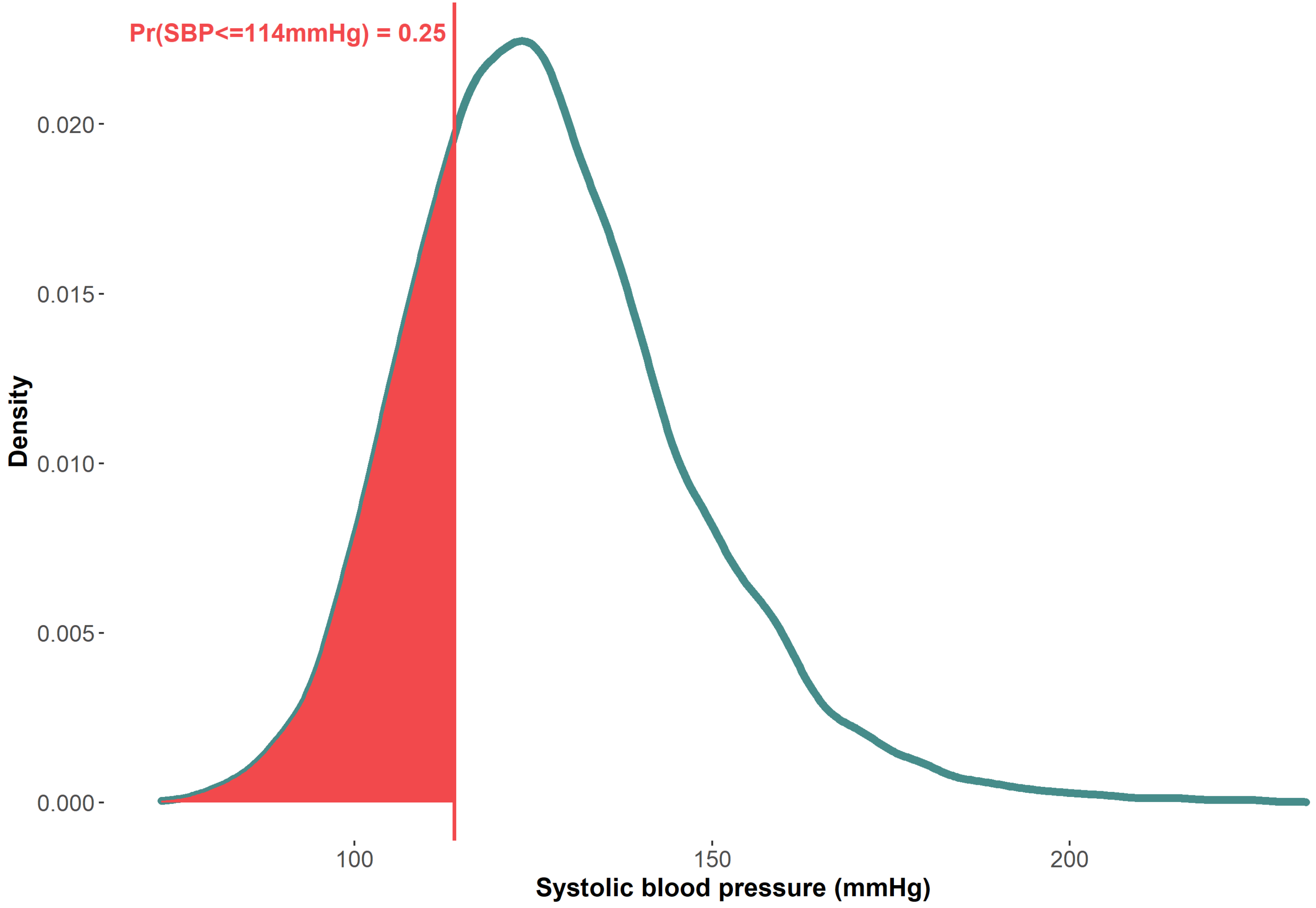
The Law of Iterated Expectations allows us to determine the mean of the unconditional distribution using information on the means of the conditional distribution

50^{th} quantile is the value of SBP such that 50% of values lie below it

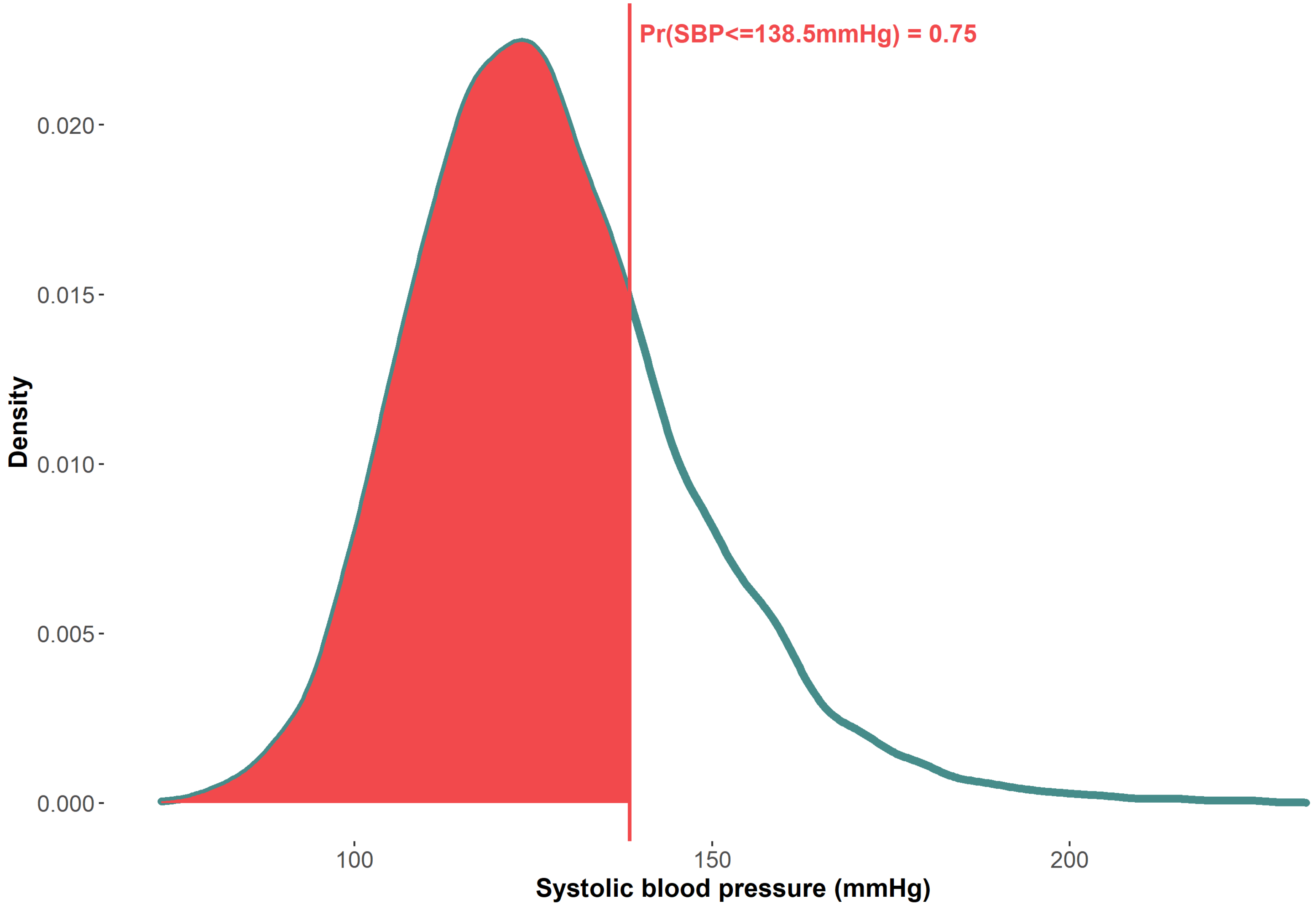


50^{th} quantile = median

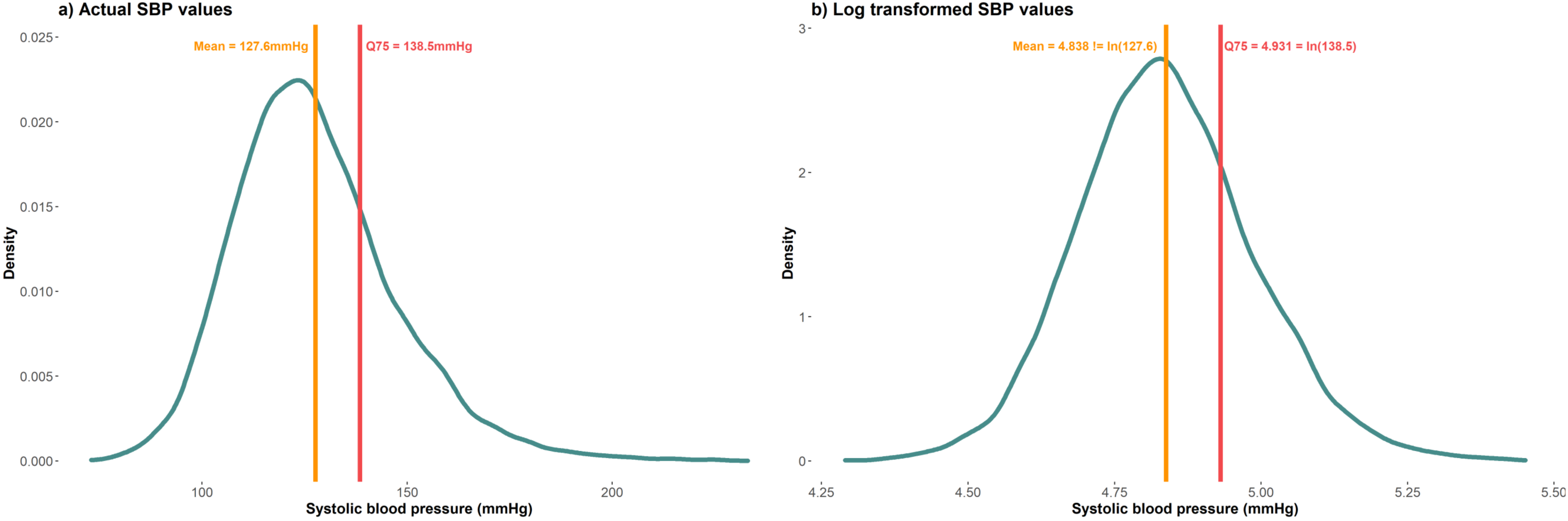
25th quantile is the value of SBP such that 25% of values lie below it



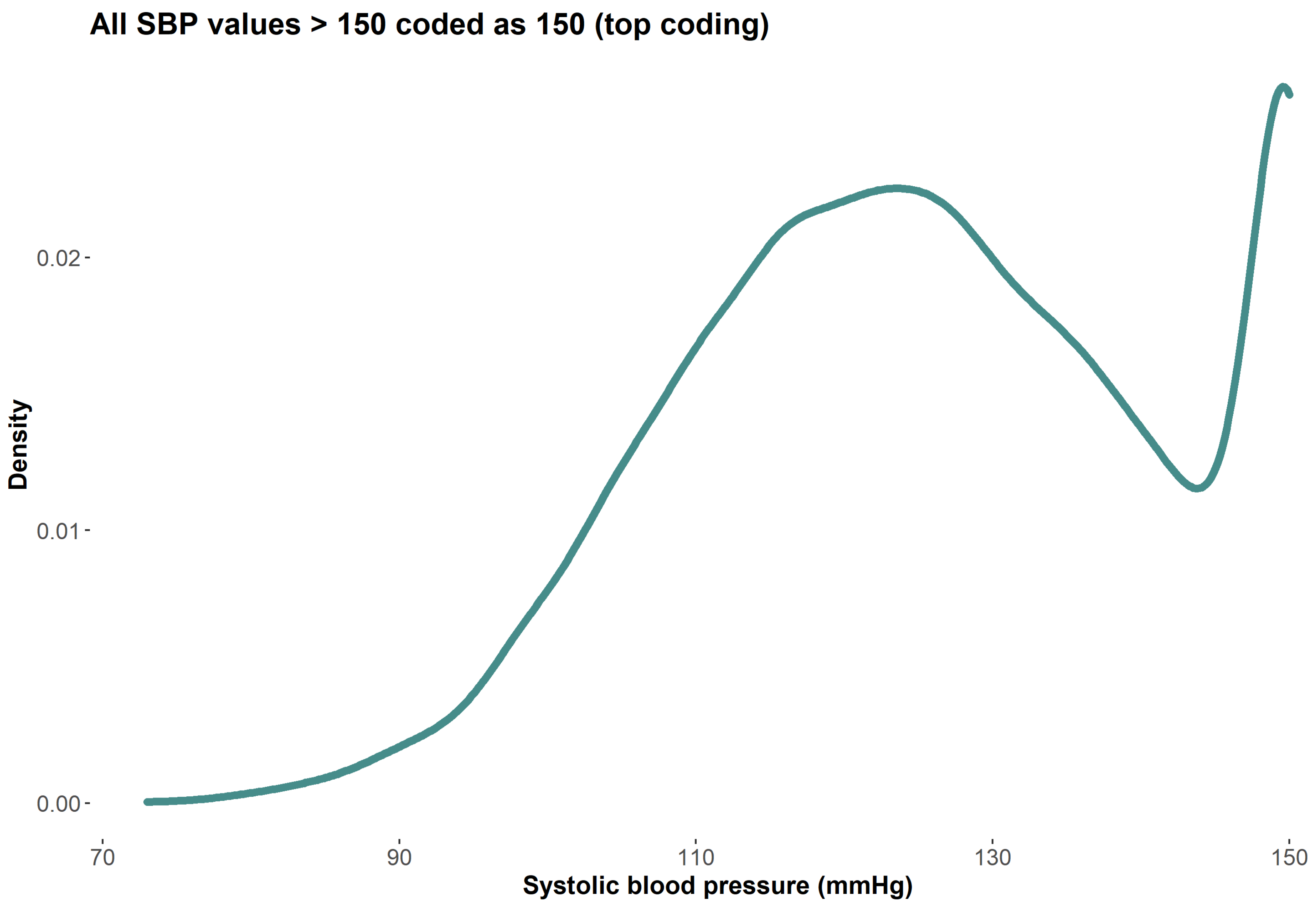
75^{th} quantile is the value of SBP such that 75% of values lie below it



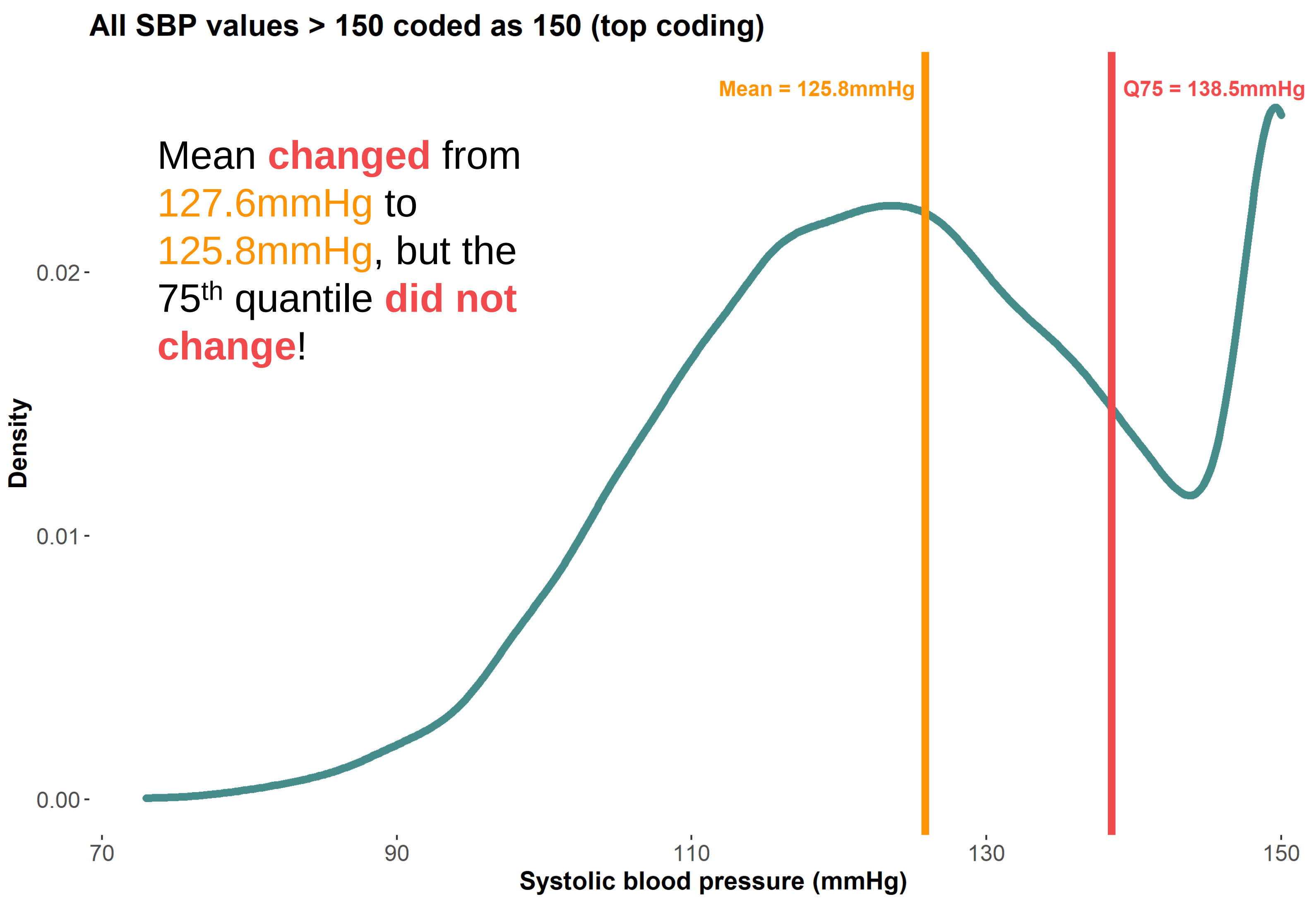
Robust quantiles: Equivariance to monotonic transformations 💪💪💪



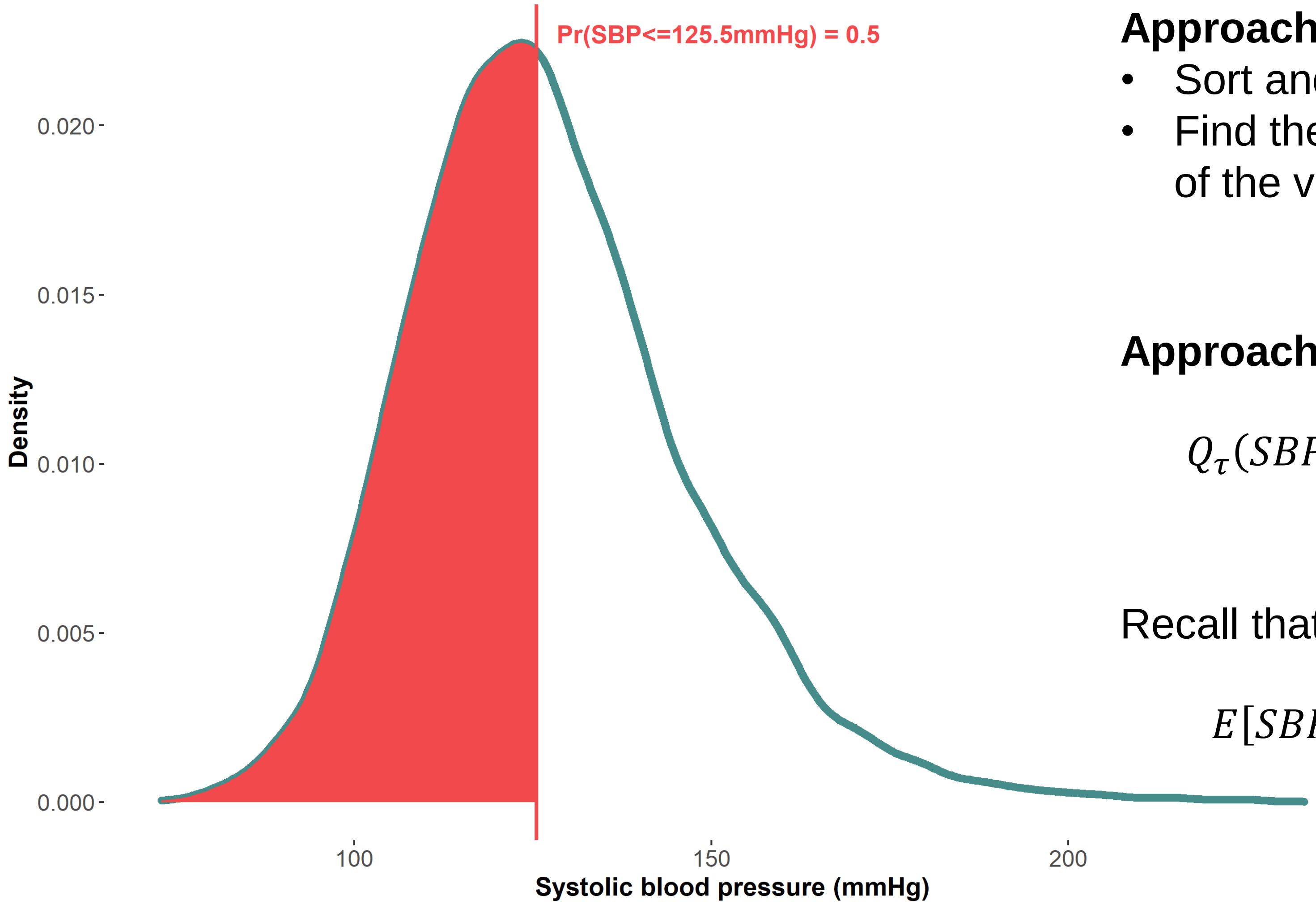
Robust quantiles: Most quantiles unaffected by top/bottom coding💪💪💪



Robust quantiles: Most quantiles unaffected by top/bottom coding 💪💪💪



How can we estimate quantiles of the unconditional SBP distribution?



Approach 1:

- Sort and arrange all values of SBP
- Find the value of SBP such that, say, 50% of the values lie below it

Approach 2:

$$Q_\tau(SBP) = \boxed{\min_a \sum_{i=1}^{N} \rho_\tau(SBP_i - a)}$$

Recall that for the mean,

$$E[SBP] = \boxed{\min_a \sum_{i=1}^{N} (SBP_i - a)^2}$$

Rho rho rho your boat

$$\rho_\tau(u) = u\big(\tau - \boldsymbol{I}(u < 0)\big)$$

Rho rho rho your boat

$$\rho_\tau(u) = u\big(\tau - \boldsymbol{I}(u < 0)\big)$$

$$\boldsymbol{I}(u < 0) = \begin{cases} 1, & u < 0 \\ 0, & u \geq 0 \end{cases}$$

Rho rho rho your boat

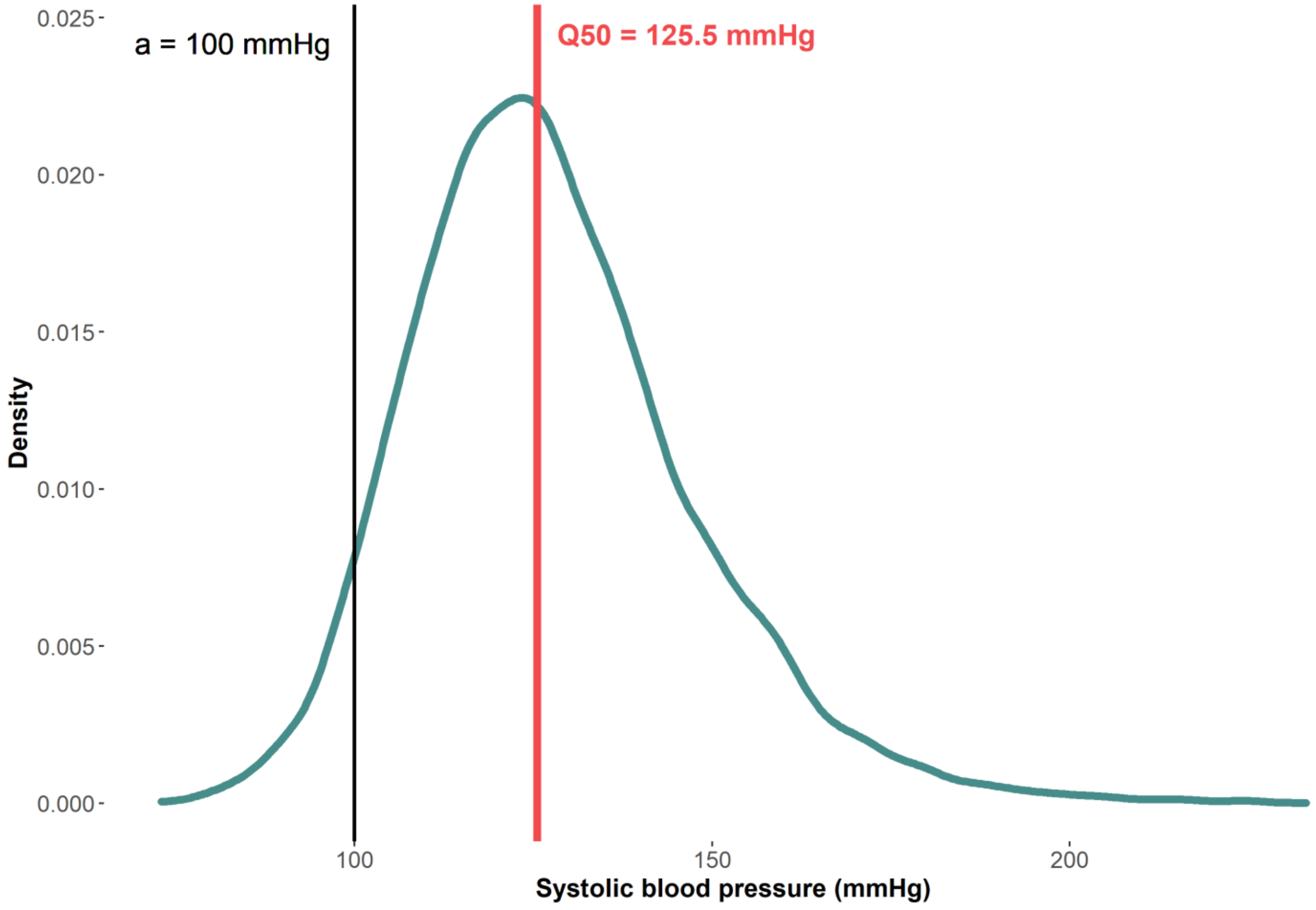
$$\rho_\tau(u) = u\big(\tau - \boldsymbol{I}(u < 0)\big)$$

$$\boldsymbol{I}(u < 0) = \begin{cases} 1, & u < 0 \\ 0, & u \geq 0 \end{cases}$$

$$\rho_\tau(u) = \begin{cases} (\tau - 1)u, & u < 0 \\ \tau u, & u \geq 0 \end{cases}$$

$$= \boxed{(1 - \tau)\boldsymbol{I}(u < 0)|u|} + \boxed{\tau \boldsymbol{I}(u \geq 0)|u|}$$

Visualizing $\rho_{0.5}(SBP_i - a)$



Suppose $\tau = 0.5$ and $a = 100$. How can we plot $\rho_{0.5}(SBP_i - 100)$?

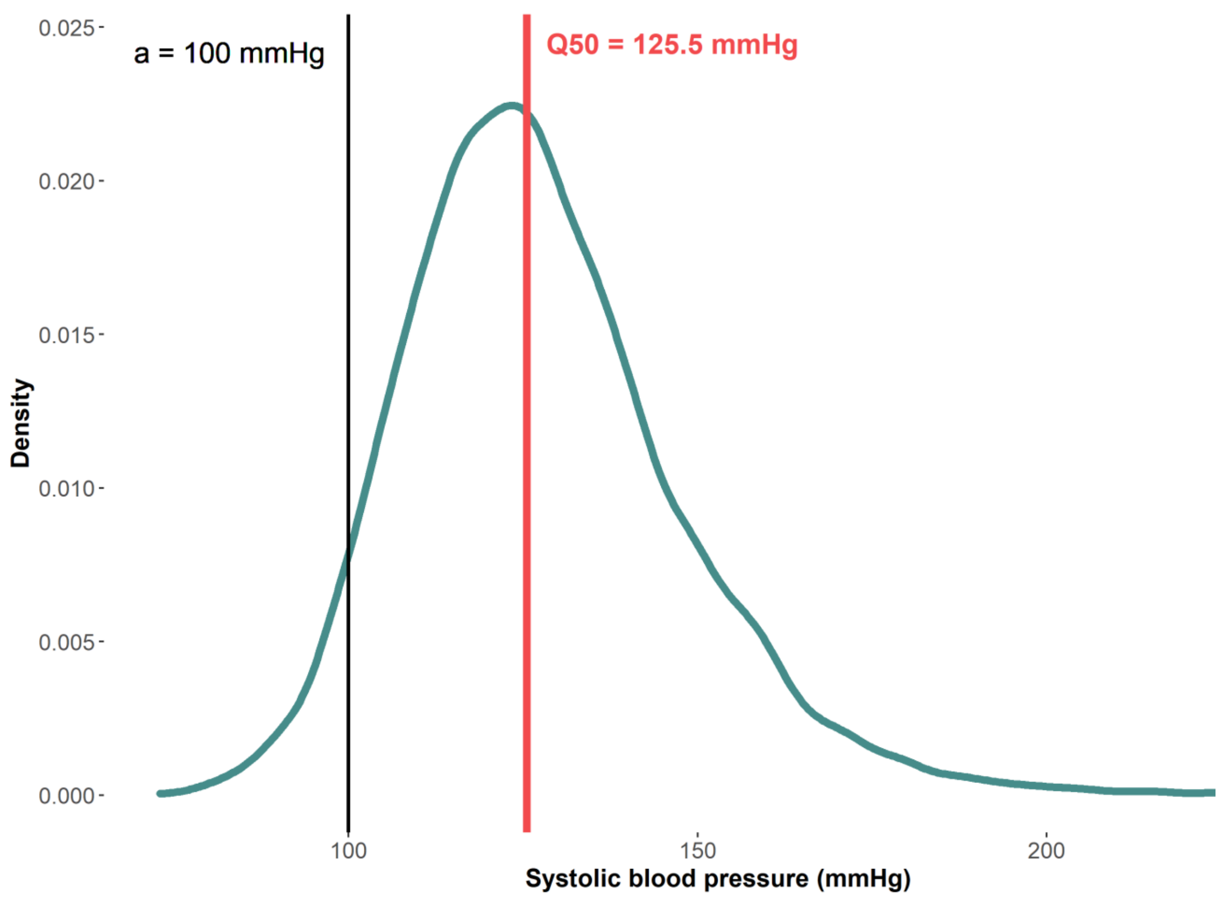
Step 1: Notice that

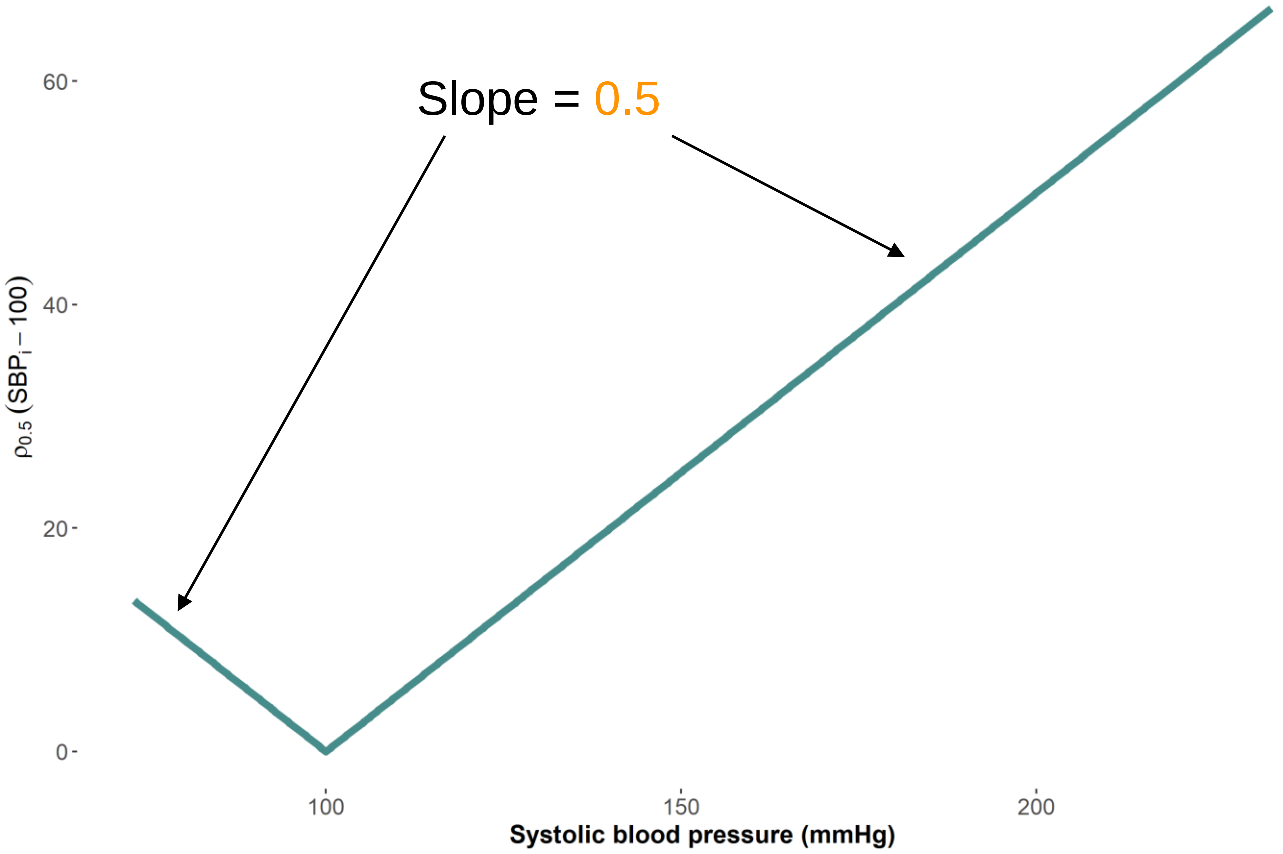
$$\rho_{0.5}(SBP_i - 100)$$
$$= 0.5\boldsymbol{I}(SBP_i - 100 < 0)|SBP_i - 100|$$
$$+ 0.5\boldsymbol{I}(SBP_i - 100 \geq 0)|SBP_i - 100|$$

Step 2: Find all values of SBP such that $SBP_i - 100 < 0$. When $SBP_i - 100 < 0$, $\rho_{0.5} = 0.5 * |SBP_i - 100|$. Now, calculate $\rho_{0.5}(SBP_i - 100)$ when $SBP_i - 100 < 0$.

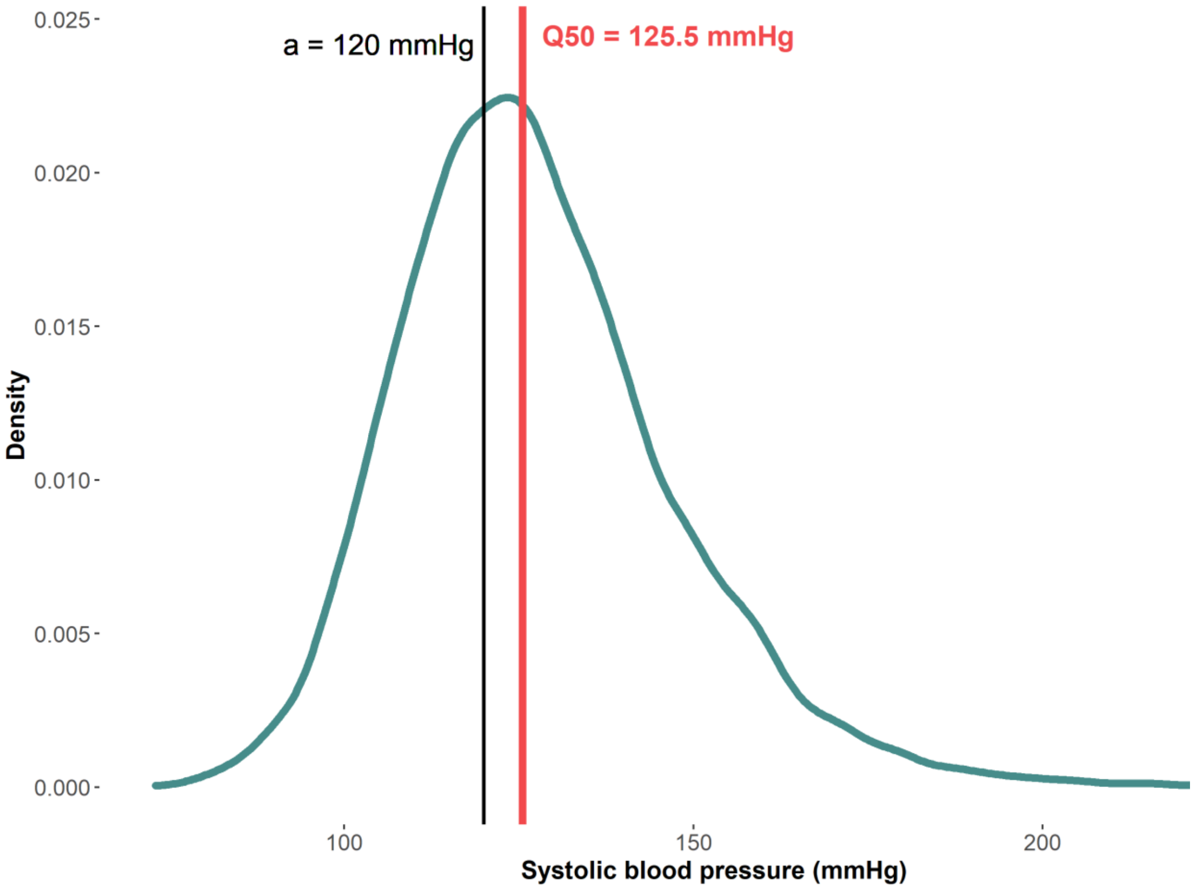
Step 3: Repeat step 2 for $SBP_i - 100 \geq 0$. Then have fun plotting the calculated points!

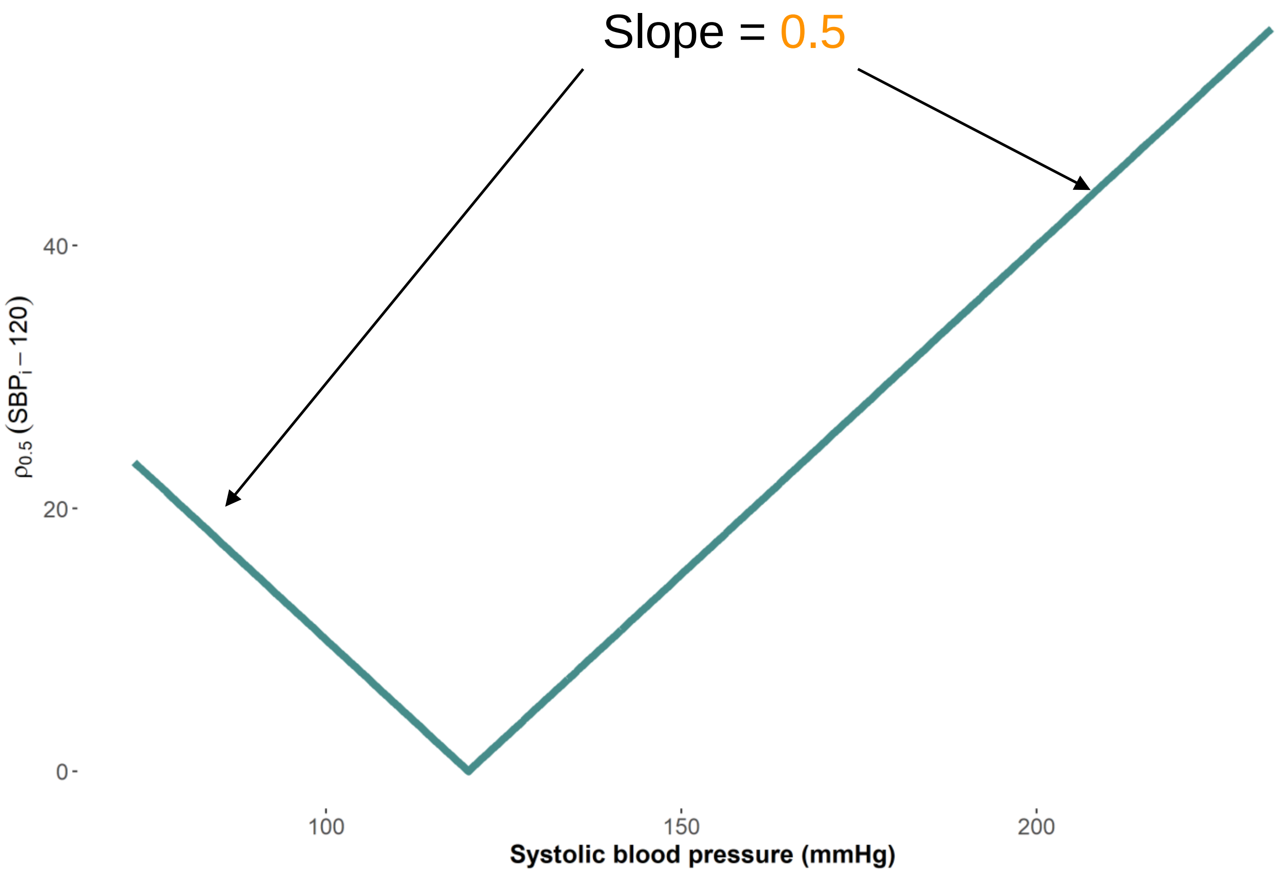
Choose $a = 100$. Now let's plot $\rho_{0.5}(SBP_i - 100)$





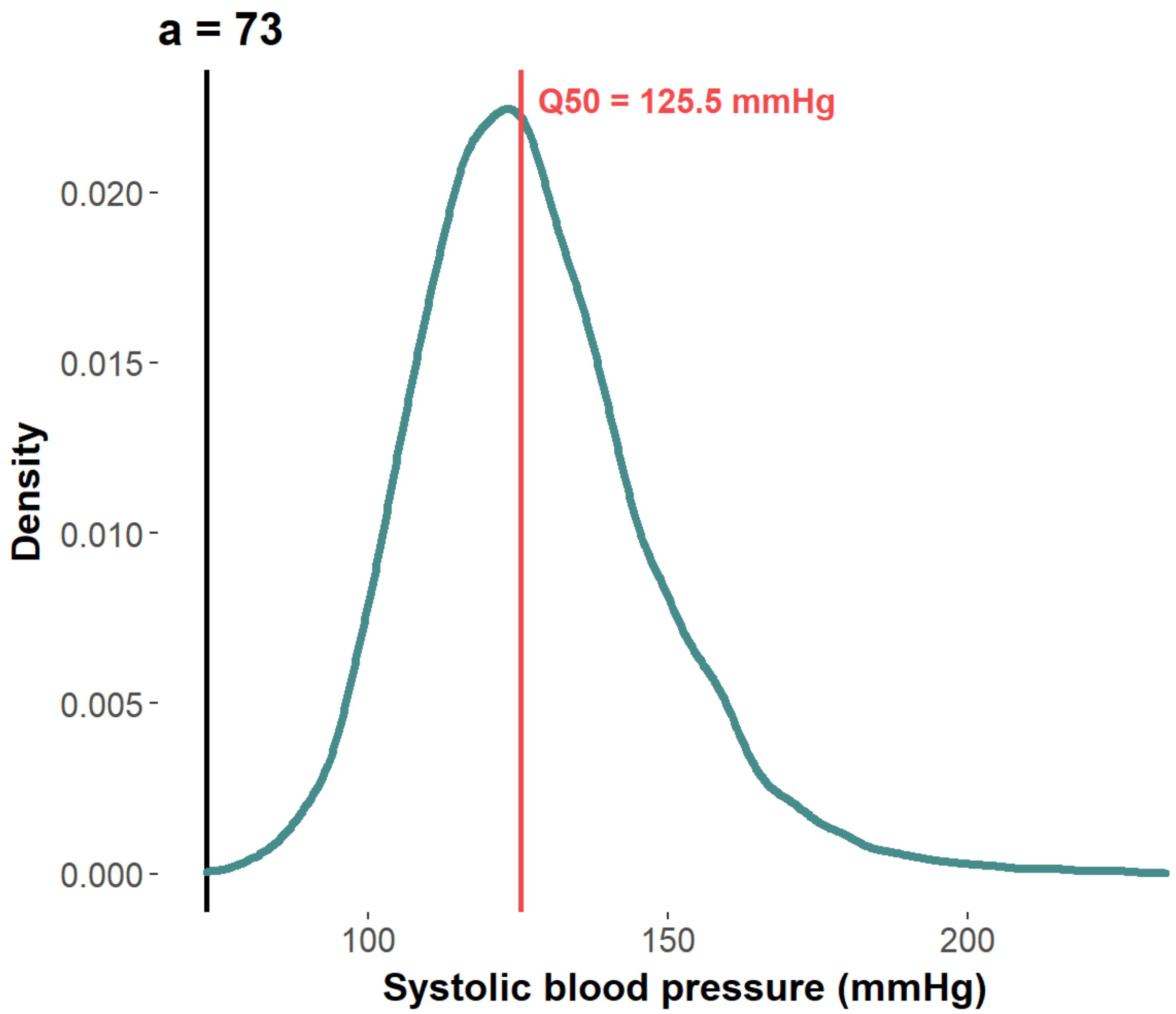
Choose $a = 120$. Now let's plot $\rho_{0.5}(SBP_i - 120)$

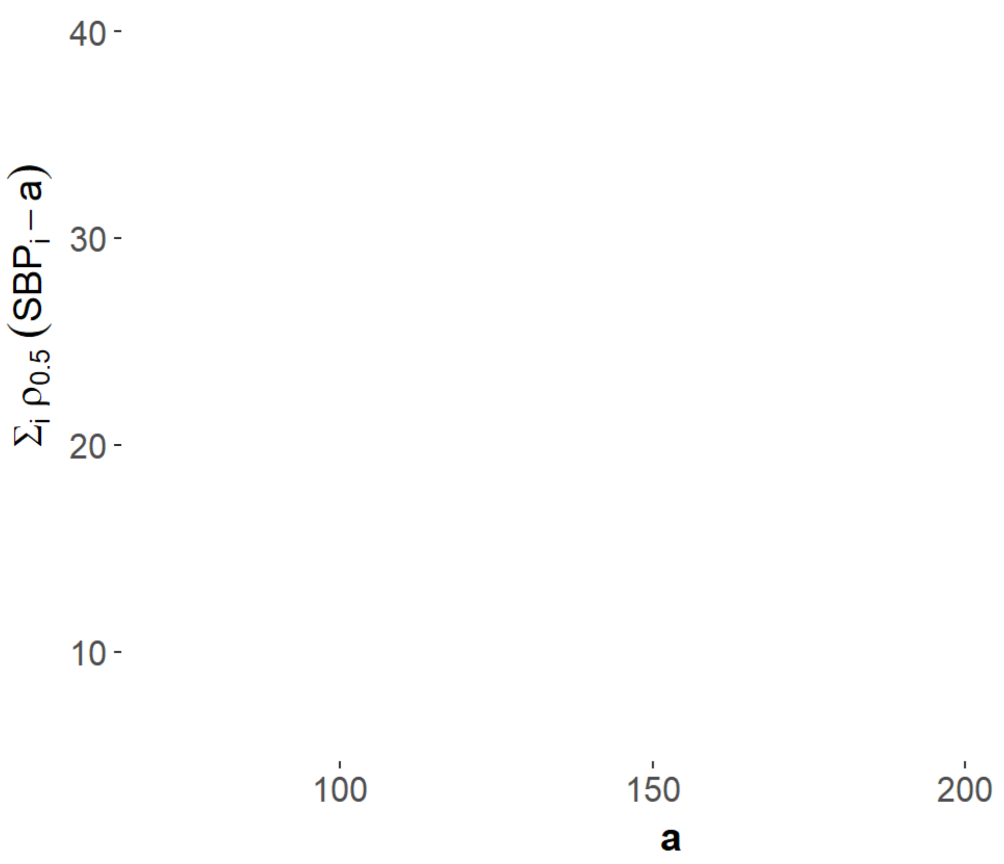




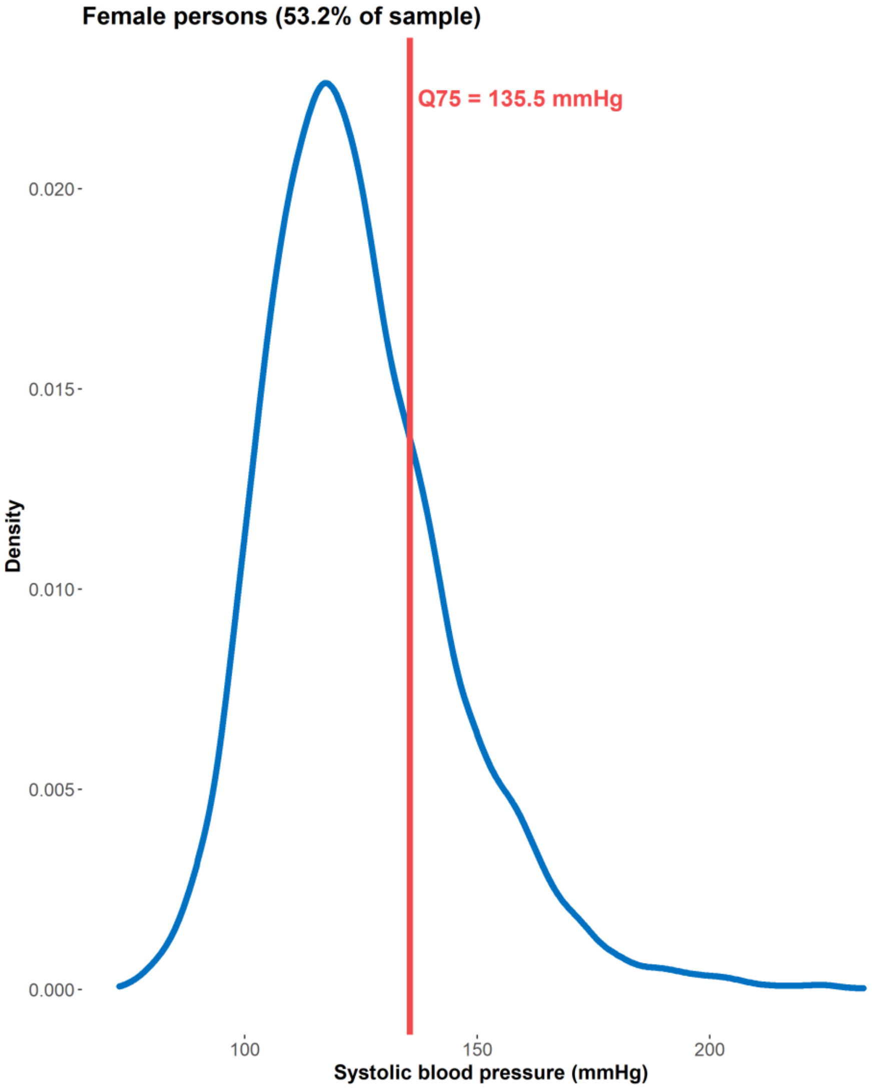
Visualizing the minimization for the median

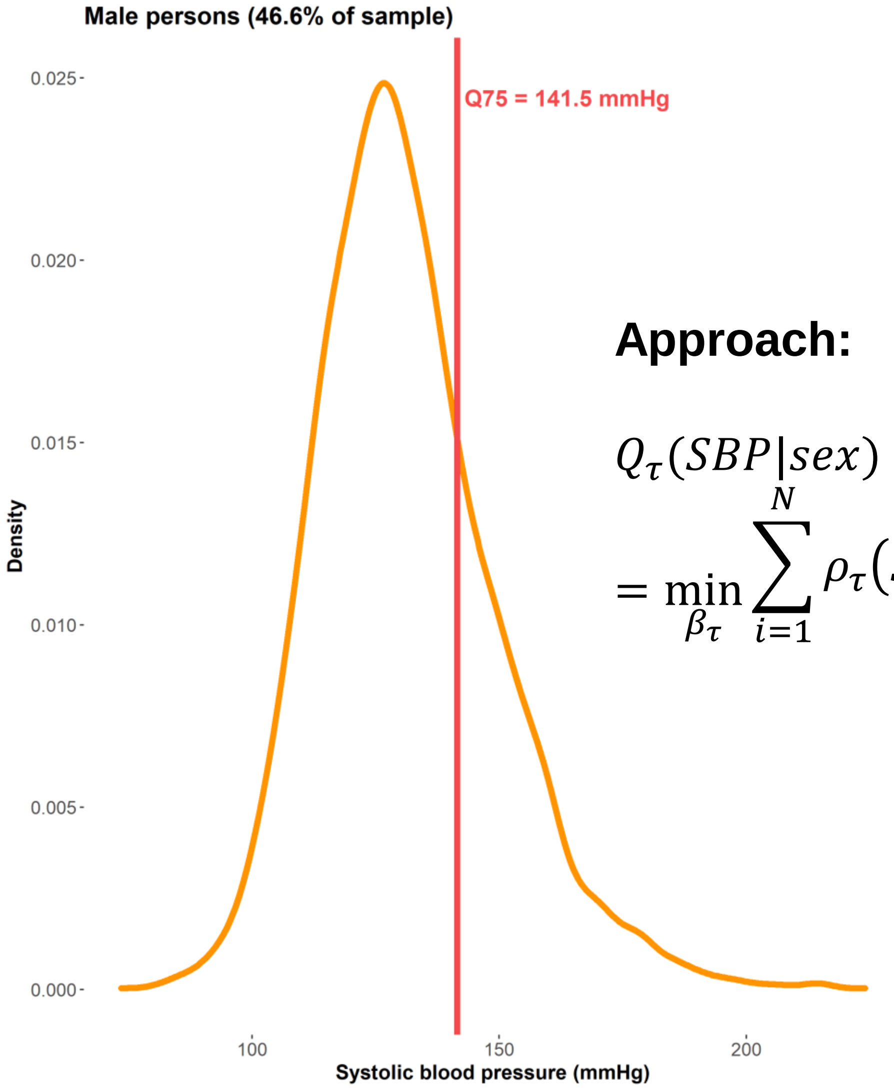
$$\min_{a} 0.5 \sum_{i=1}^{N} |SBP_i - a|$$





Quantiles of the SBP distribution conditional on sex

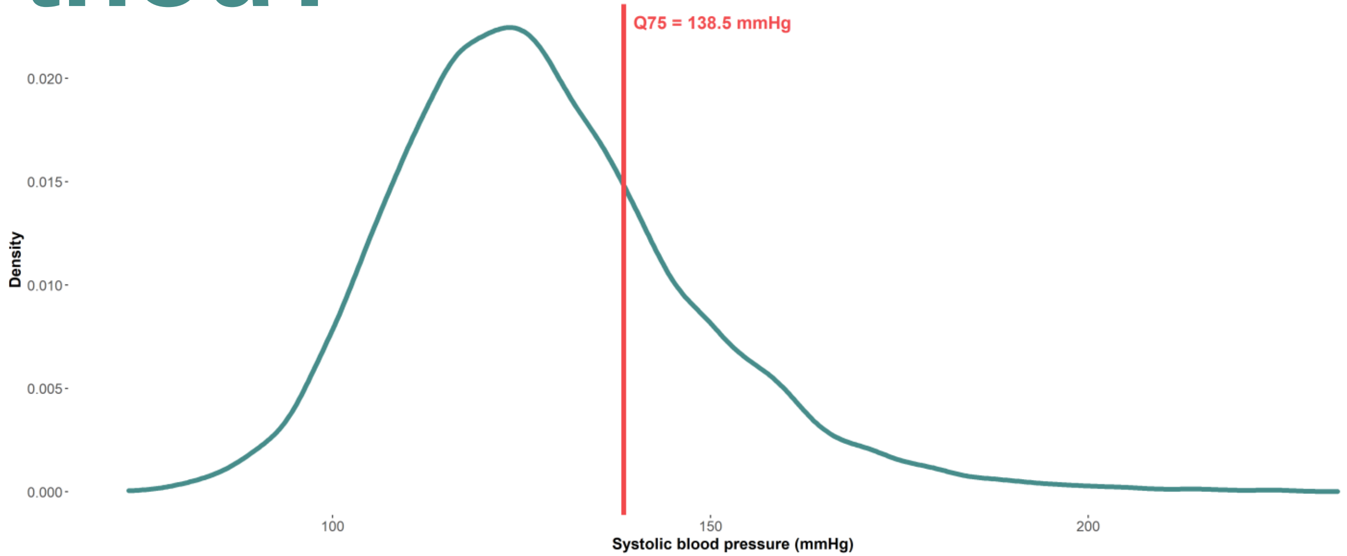


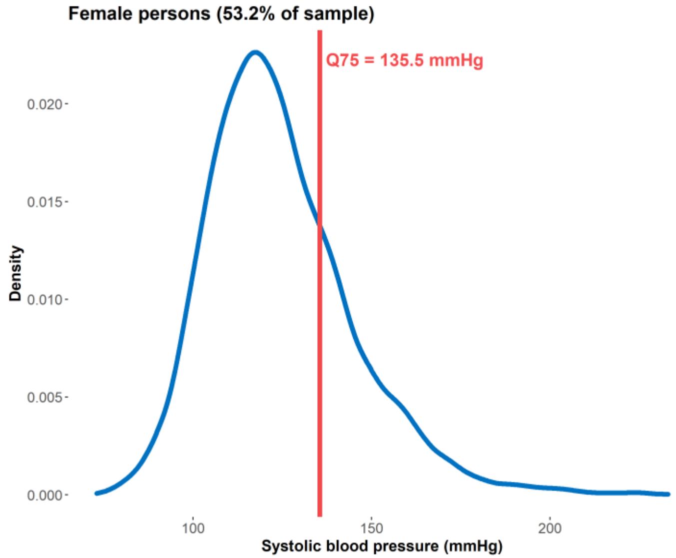


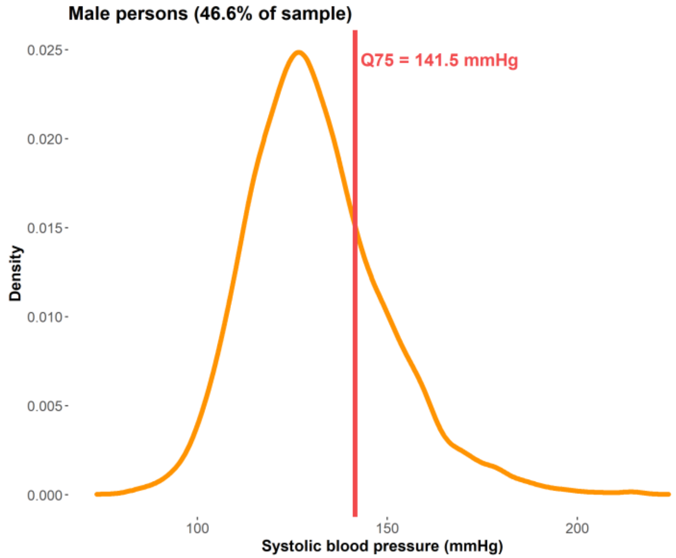
Approach:

$$Q_\tau(SBP|sex) = \min_{\beta_\tau} \sum_{i=1}^{N} \rho_\tau\big(SBP_i - g(sex_i, \beta_\tau)\big)$$

Law of Iterated Quantiles, where art thou?







The Law of Iterated Quantiles does not exist (yet)

Note that:

$$0.532 * 135.5 + 0.466 * 141.5 = 138\,mmHg$$

Which is close to the Q75 value in the unconditional distribution but not exact

The lack of a Law of Iterated Quantiles means that we usually cannot use information about quantiles in the conditional distribution to learn about the same quantiles in the unconditional distribution

Key takeaways: Means

1. We can calculate means by minimizing a quadratic loss function
   - Unconditional mean: $\min_{a} \sum_{i=1}^{N} (y_i - a)^2$
   - Conditional mean: $\min_{\beta} \sum_{i=1}^{N} \left(y_i - g(x_i, \beta)\right)^2$

2. The Law of Iterated Expectations allows us to learn about the unconditional mean from the conditional means

Key takeaways: Quantiles

1. The τ^{th} quantile is the value of a random variable such that τ% of the random variable lies below that value

2. We can calculate quantiles by minimizing the check function
   - Unconditional quantile: $\min_{a} \sum_{i=1}^{N} \rho_\tau(y_i - a)$
   - Conditional quantile: $\min_{\beta_\tau} \sum_{i=1}^{N} \rho_\tau\left(y_i - g(x_i, \beta_\tau)\right)$

3. There is no "Law of Iterated Quantile" (yet)
   - Need to be clear about whether unconditional or conditional quantiles are of interest

10-minute break

The tyranny of *l'homme moyen*

i.e., a review of linear regression

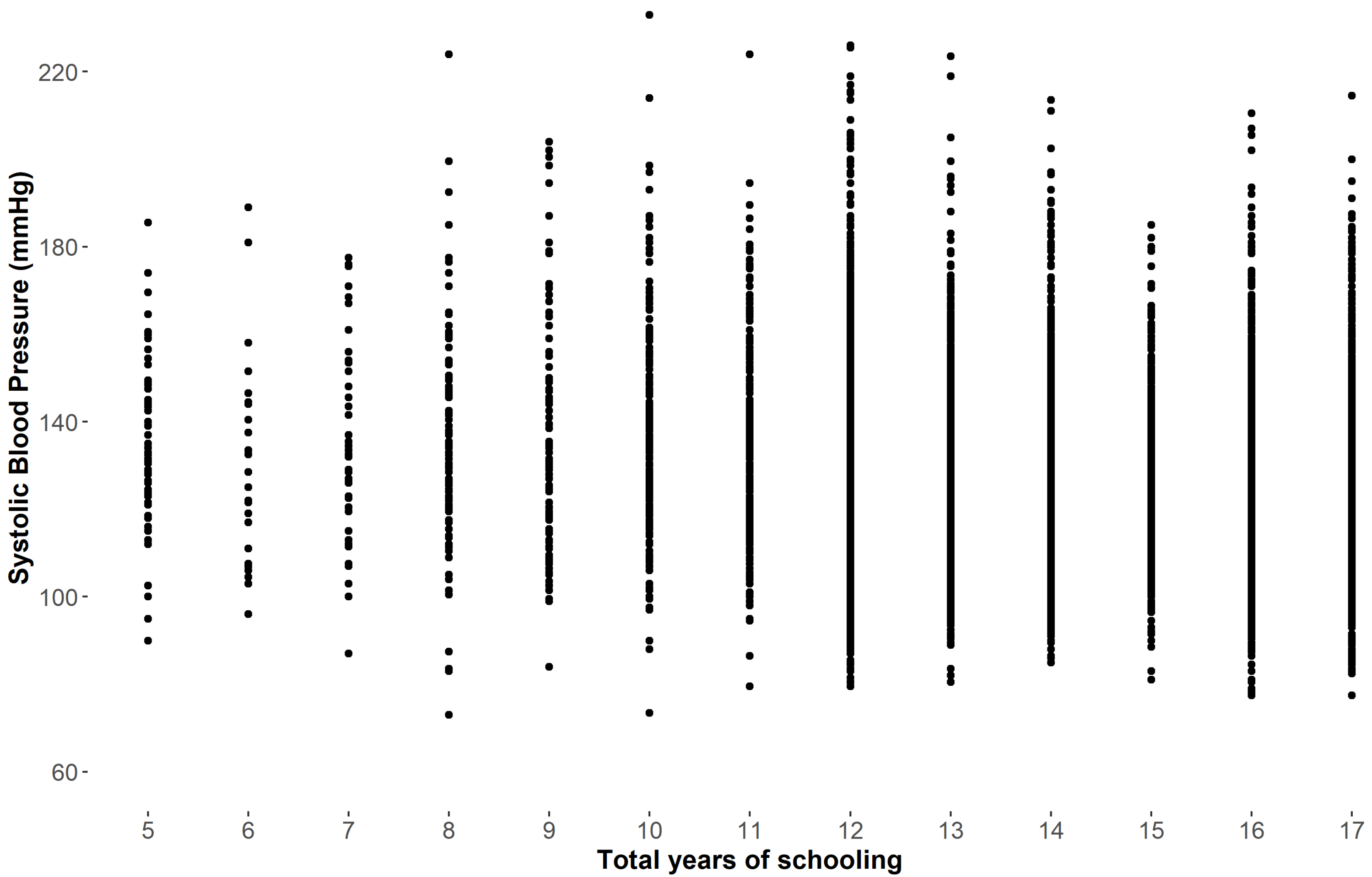
Learning aims

1. Linear regression and the conditional expectation function

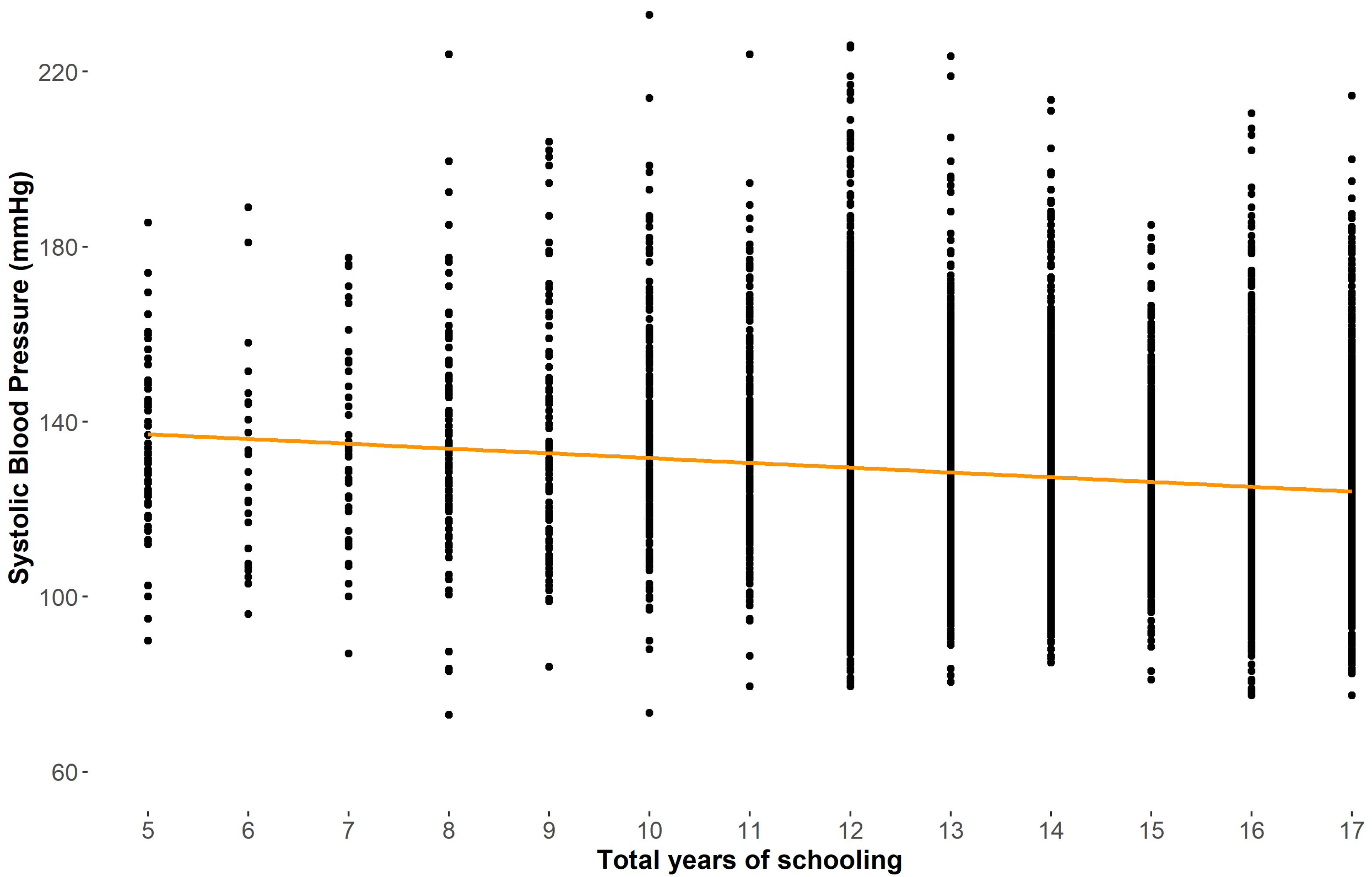
2. Estimating coefficients and standard errors in linear regression

3. Interpreting linear regression results

Objective: Quantify the association of education with SBP



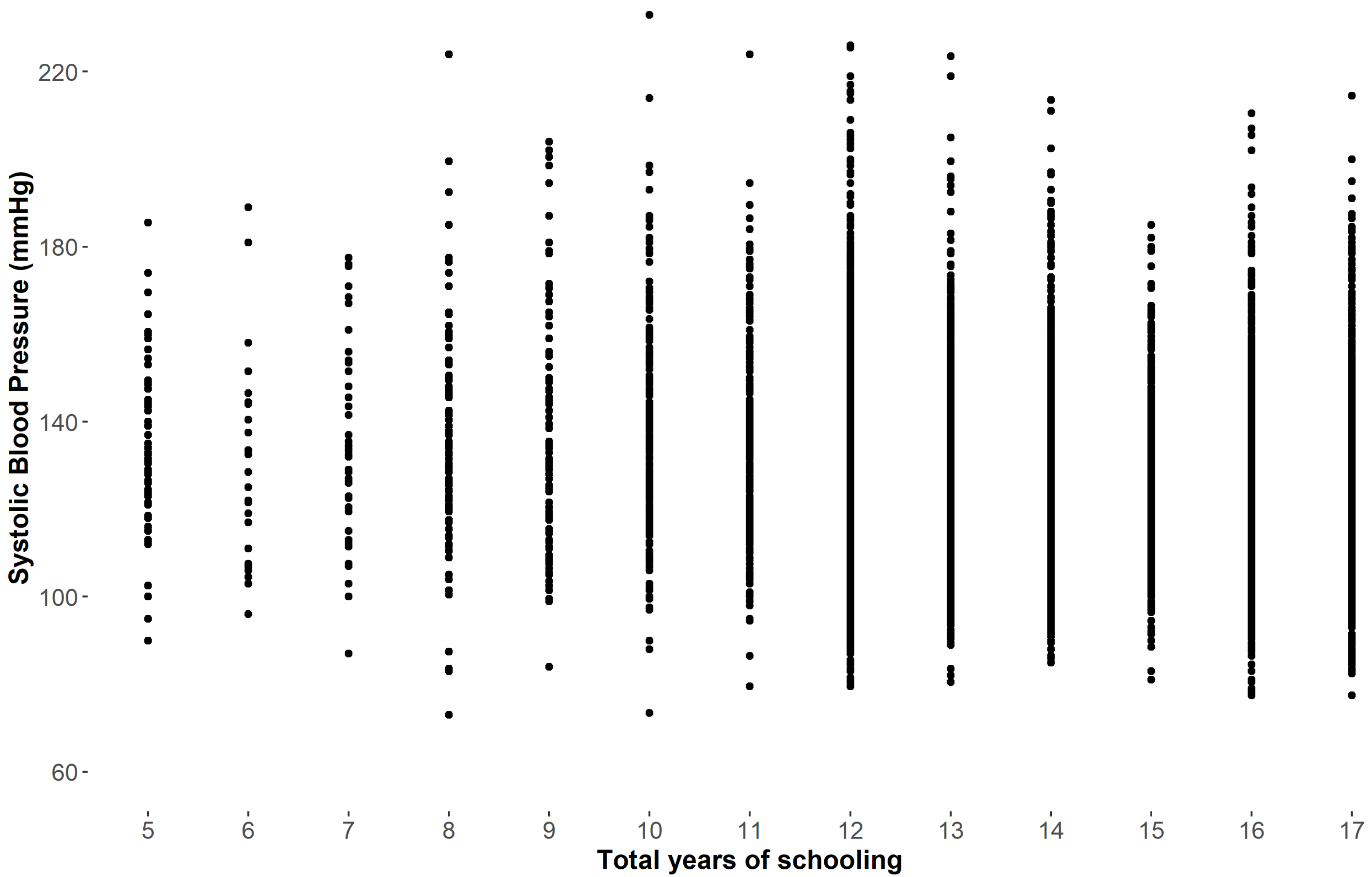
Potential solution:
`lm(sbp ~ schlyrs)`



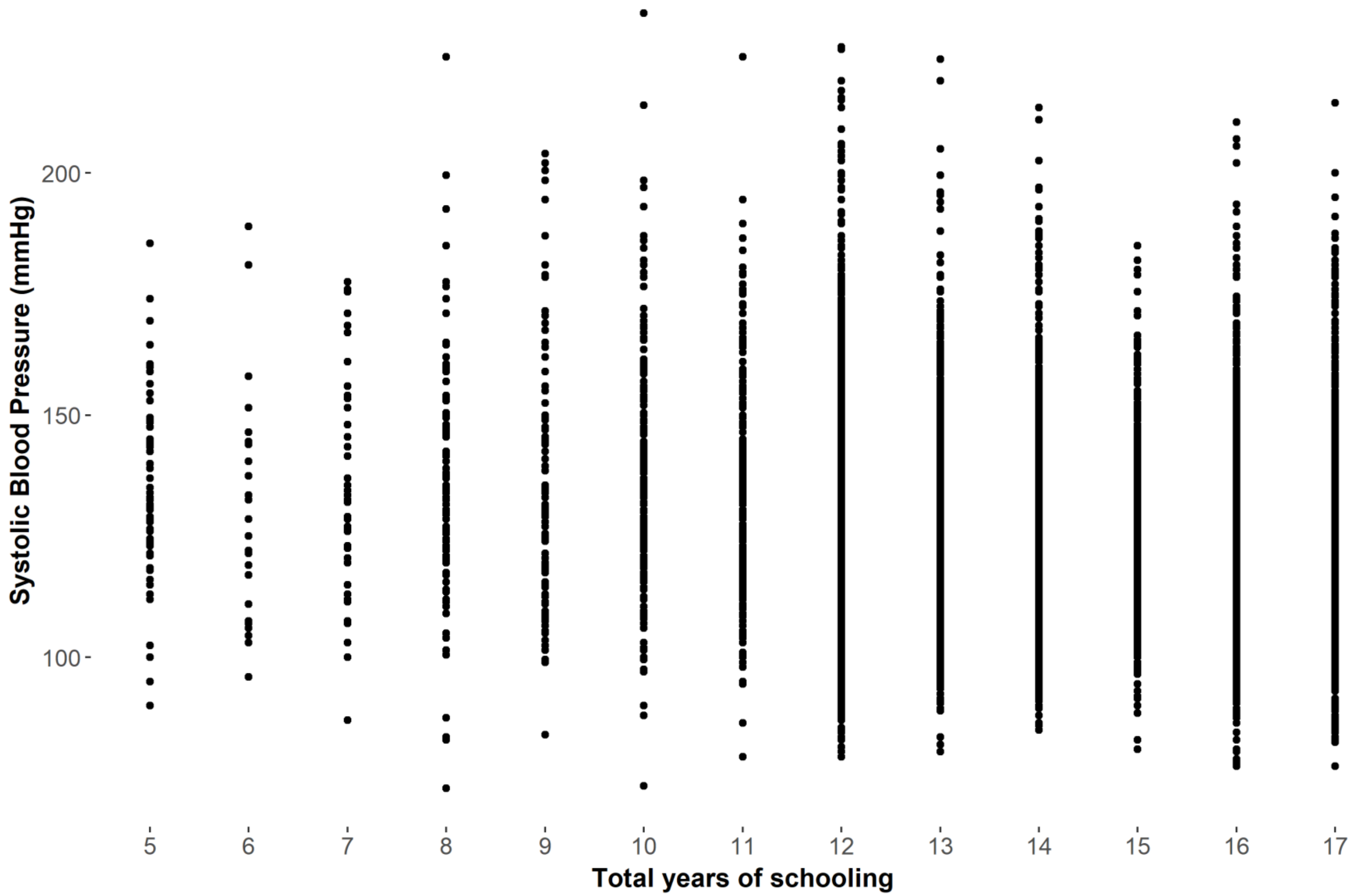
But what exactly is linear regression modeling?


Bernie
I am once again asking
you to suspend your disbelief
imgflip.com

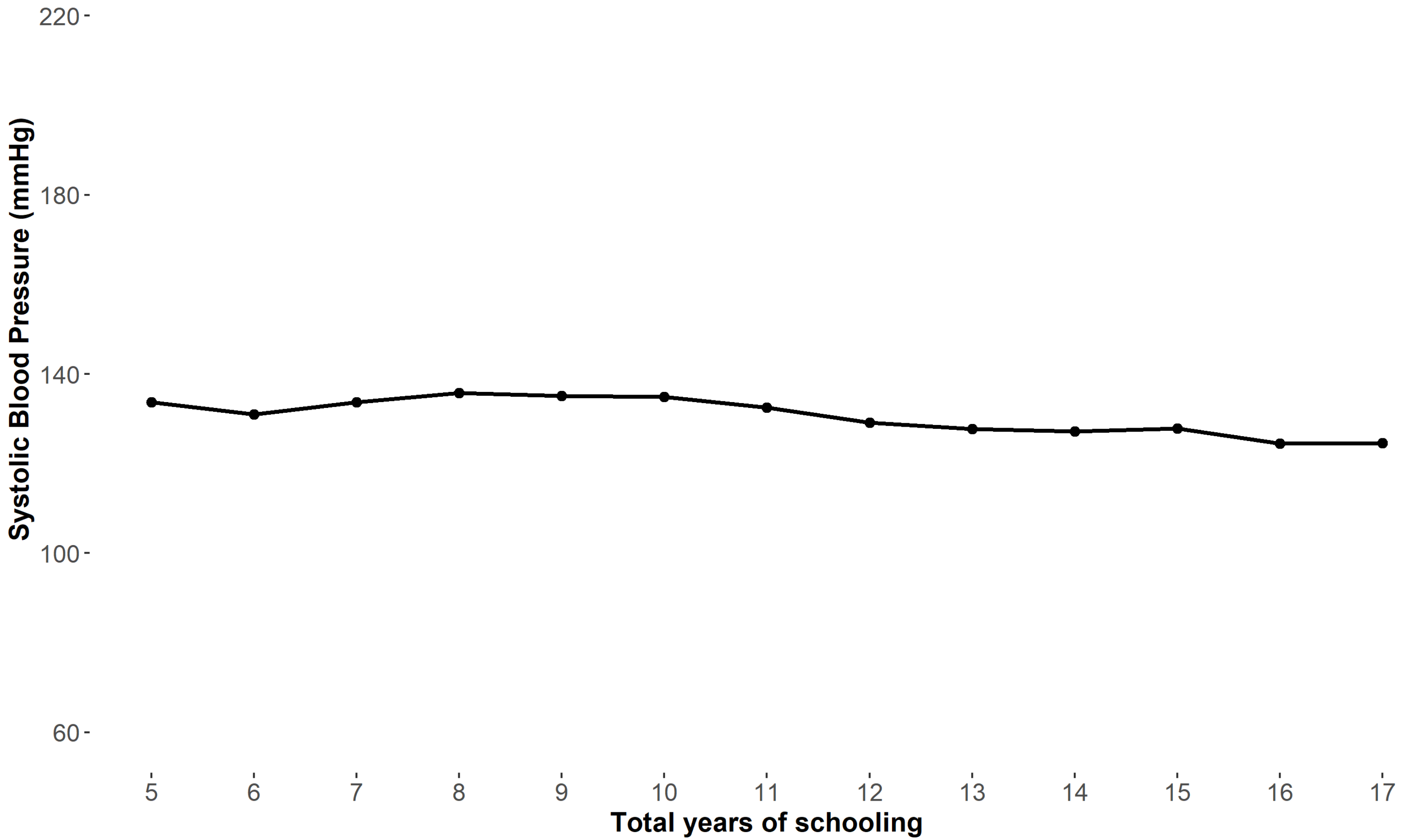
Imagine that this figure presents SBP data for the whole population



Calculate mean SBP by education level (i.e., the conditional expectation)

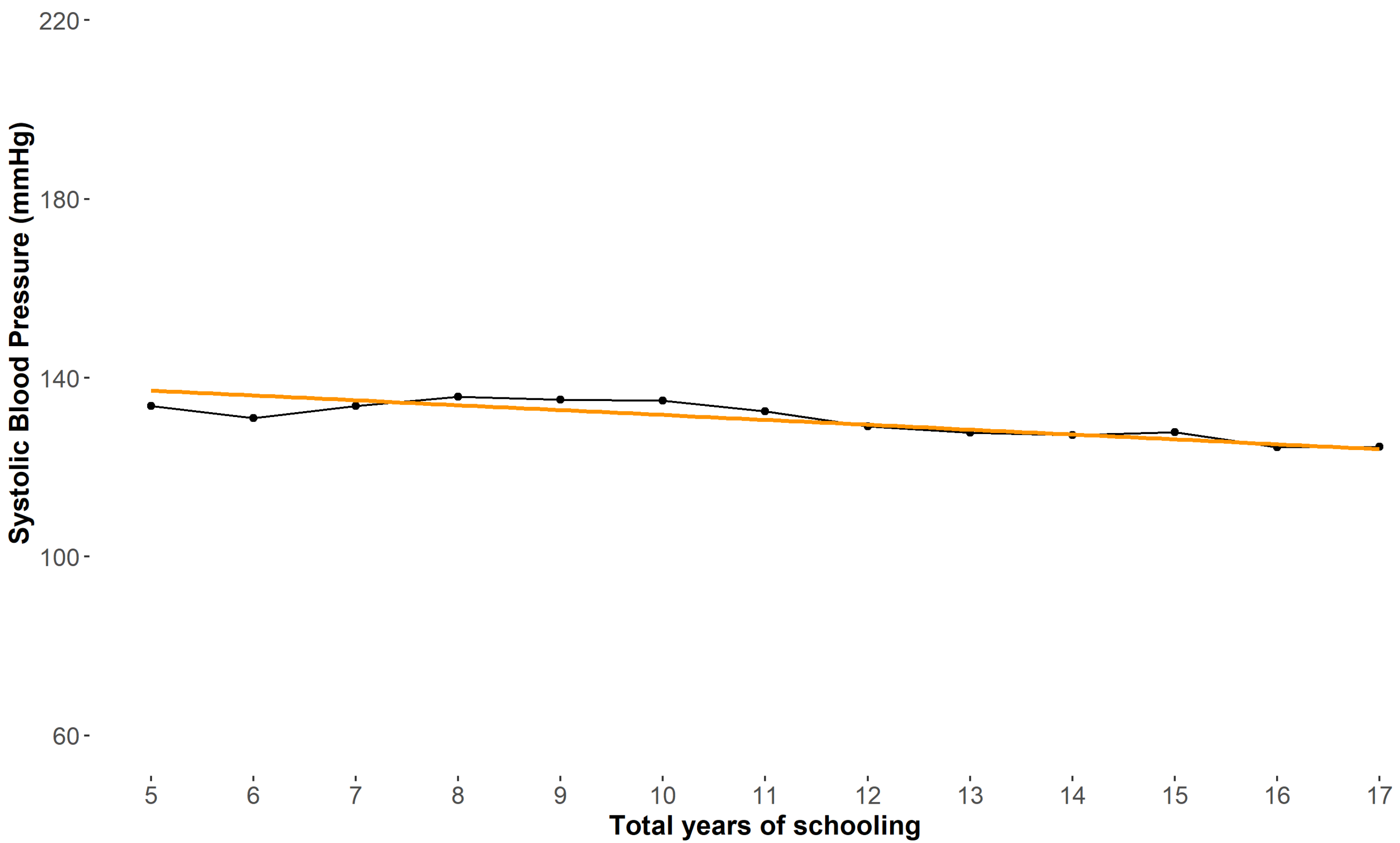


Conditional expectation function (CEF) links the conditional means

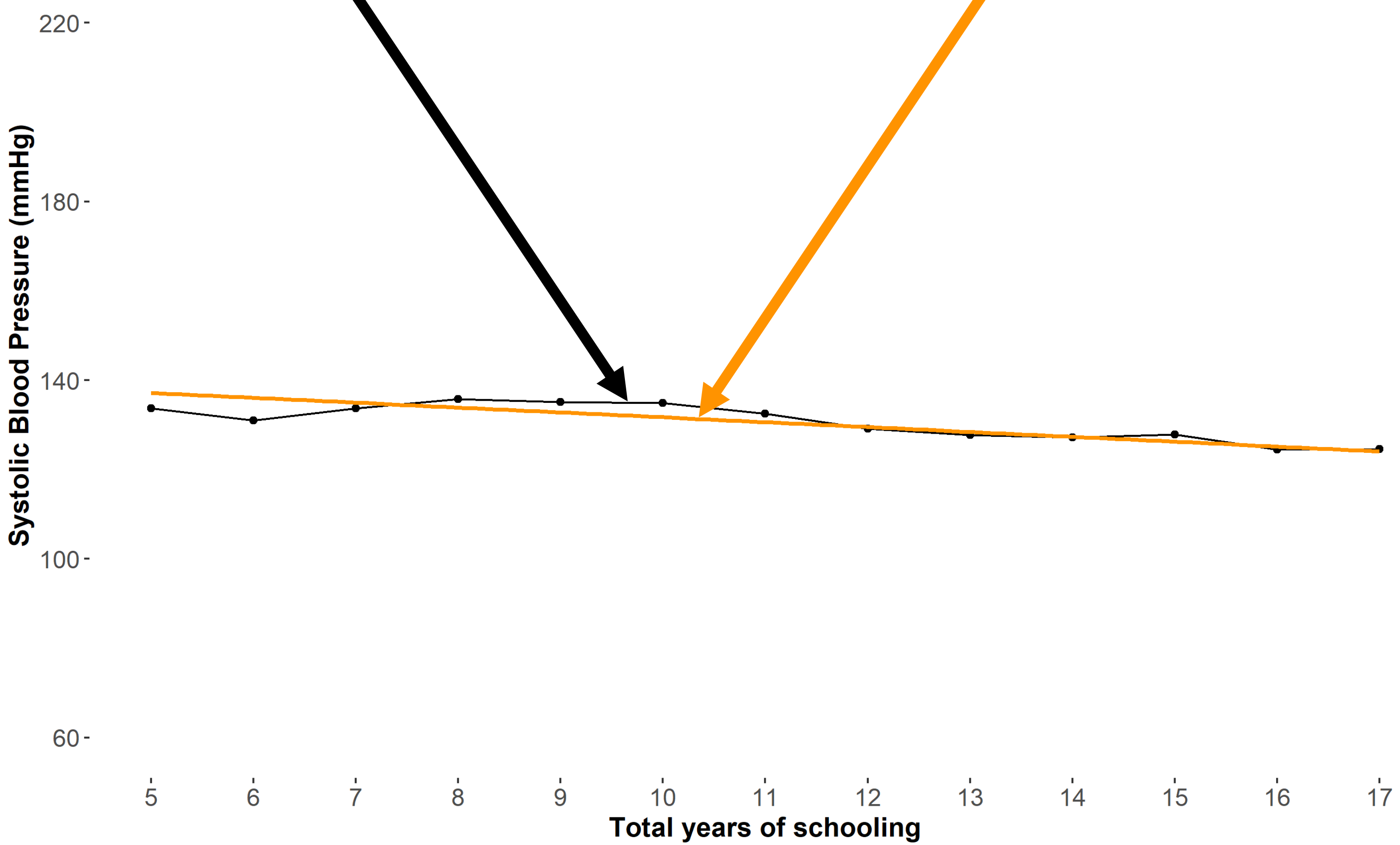


The Conditional Expectation Function tells us how the conditional mean of the outcome changes as we change values of conditioning variable

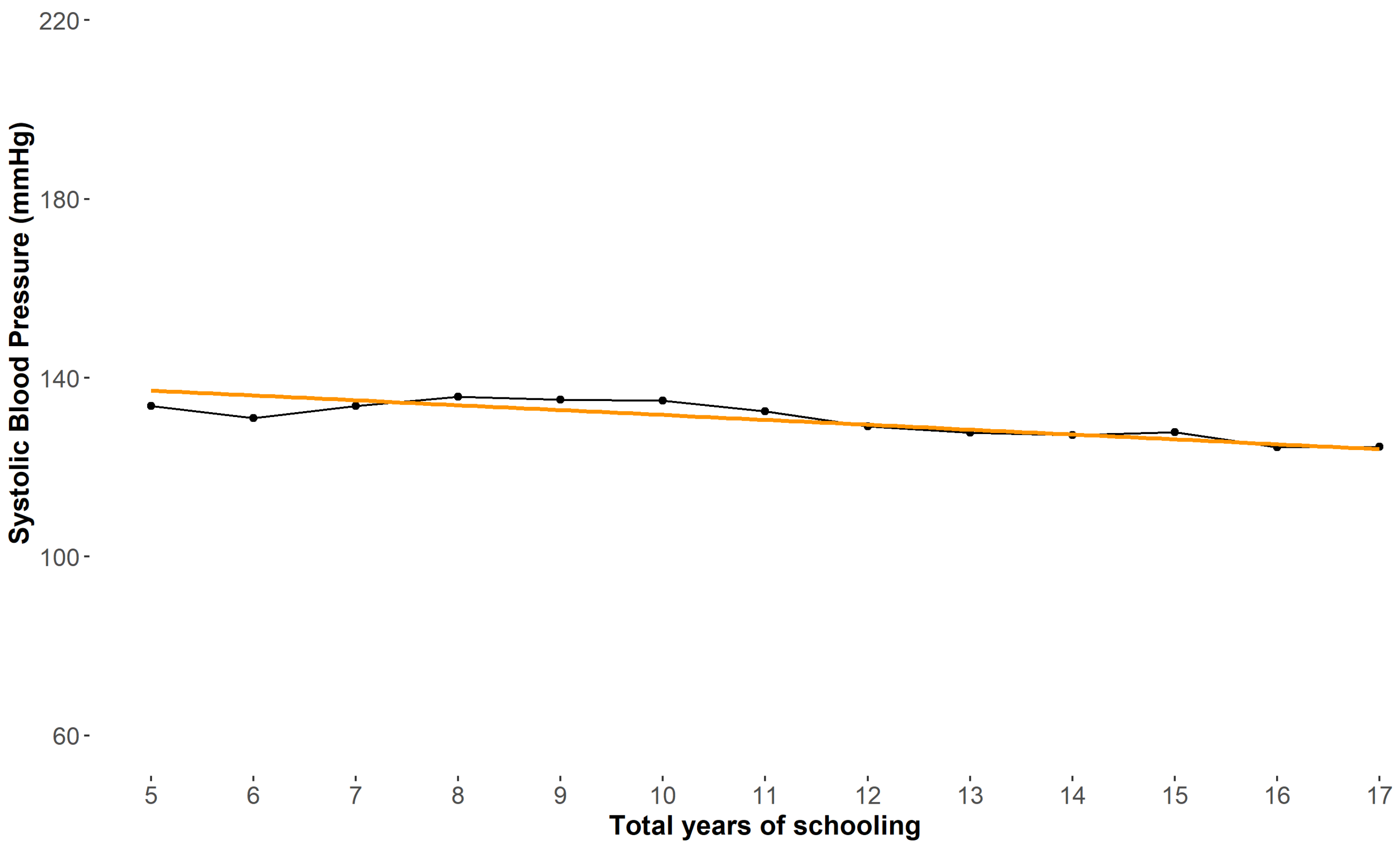
Population regression line ≈ CEF



$$\boxed{E[SBP|schlyrs]} = \beta_0 + \boxed{\beta_1} schlyrs_i$$



Linear regression in sample ≈ population regression line ≈ CEF



Since it is a model of the conditional expectation function, linear regression provides us with an estimate for how the mean of the outcome changes as we change the exposure by one unit

In the context of our example, linear regression answers the question: By how much does mean SBP change for each additional year of schooling?

How do we estimate the coefficients of a linear regression?

Our model of the world

- Imagine that our model of the world (i.e., the data generating process or DGP) is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\therefore \epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

- So, in the case of our empirical example, our model is:

$$SBP_i = \beta_0 + \beta_1 schlyrs_i + \epsilon_i$$

Population regression line ≈ CEF

- Recall that the population regression line is a model of the CEF

$$E[Y|X] = \beta_0 + \beta_1 x_i$$

- Here, we make two key assumptions about the error

$$E[\epsilon|X] = 0 \Rightarrow E[\epsilon] = 0$$

$$Var(\epsilon|X) = Var(\epsilon) = \sigma^2$$

Coefficients of the population regression line

- We can calculate the coefficients of the population regression line by minimizing the sum of squared errors

$$(\beta_0, \beta_1) = \min \sum_i \epsilon_i^2 = \min_{\beta_0, \beta_1} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

- Notice that this is how we find the conditional expectation!

$$E[Y|X] = \min_{\beta} \sum_i \big(y_i - g(x_i, \beta)\big)^2$$

We don’t have the population, only a sample

- In our sample, our regression of interest is

$$E\widehat{[Y|X]} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

Where the $\widehat{\ }$ (“hat”) indicates that we are estimating the population quantities

Coefficients of the sample linear regression

- There are several ways to estimate coefficients of the sample linear regression
  - Ordinary Least Squares
  - Moment estimator
  - Maximum likelihood estimator

- We will describe Ordinary Least Squares

Ordinary Least Squares (OLS)

- We could write the linear regression model as

$$\widehat{E[Y|X]} = X'\hat{\beta}$$

Where X is a matrix of independent variables and $\hat{\beta}$ is a vector of coefficients

- Minimize our estimate of the sum of squared errors, that is

$$\min_{\hat{\beta}} \sum_{i=1}^{N} \left(Y - X'\hat{\beta}\right)^2 \Rightarrow \hat{\beta} = (X'X)^{-1}(X'Y)$$

Interpreting coefficients of a simple linear regression

$$\widehat{E[Y|X]} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

- $\widehat{\beta_0}$ is the estimated average value of Y when $x_i = 0$ (i.e., the mean of the distribution $Y|x_i = 0$)

- $\widehat{\beta_1}$ is the estimated change in the conditional mean of Y for a one-unit change in X

How do we estimate standard errors in linear regression?

Standard assumption about the error variance

- Homoskedasticity: $Var(\epsilon|X) = Var(\epsilon) = \sigma^2$

- When the error term is homoscedastic, $se(\hat{\beta}) = \sigma^2(X'X)^{-1}$

- Yet, homoskedasticity is almost always violated in practice
  - Heteroskedasticity: error variance depends on the covariates
  - Correlated outcomes over time in longitudinal, survival, or time series data
  - Clustered data due to sampling strategy or treatment assignment mechanism

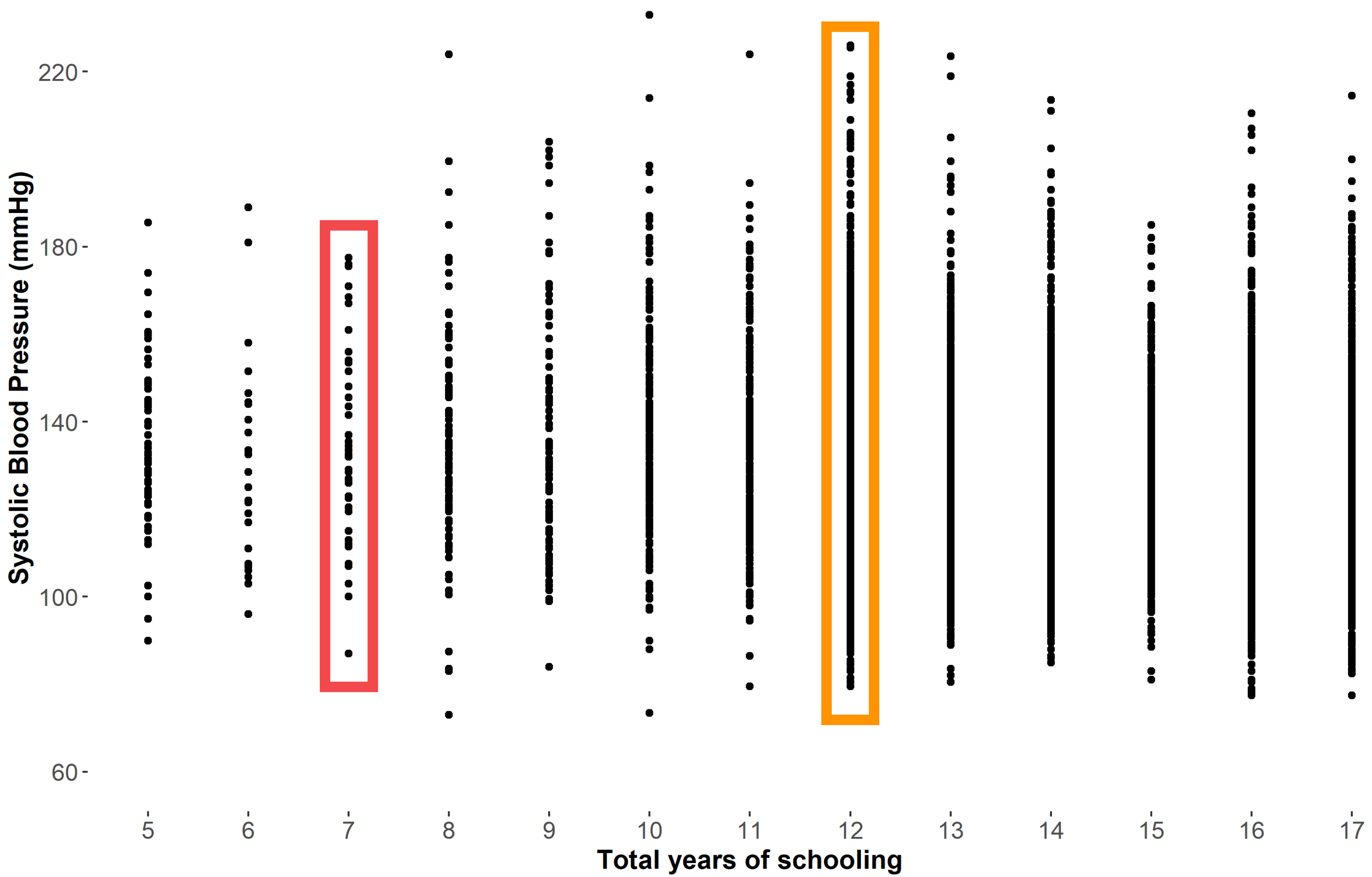
Let’s talk about heteroskedasticity

- In a heteroskedastic world,

$$Var(\epsilon|X) \neq \sigma^2$$

- The variance of the error term depends on the value of covariates included in the regression
  - That is, the errors are independent but not identically distributed

A heteroskedastic world



Robust standard errors 💪💪💪

- Estimate heteroskedasticity robust standard errors of our sample linear regression coefficients as

$$se(\hat{\beta}) = (X'X)^{-1}X'\,\Omega\,X(X'X)^{-1}$$

Where $\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} V(\epsilon_1) & 0 & \cdots & 0 \\ 0 & V(\epsilon_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & \cdots & V(\epsilon_n) \end{bmatrix}$

Robust standard errors 💪💪💪

- We can use residuals to estimate the variance-covariance matrix

$$\hat{\Omega} = \begin{bmatrix} \widehat{\epsilon_1}^2 & 0 & \cdots & 0 \\ 0 & \widehat{\epsilon_2}^2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & \cdots & \widehat{\epsilon_n}^2 \end{bmatrix} = \begin{bmatrix} (y_1 - \widehat{y_1})^2 & 0 & \cdots & 0 \\ 0 & (y_2 - \widehat{y_2})^2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & \cdots & (y_n - \widehat{y_n})^2 \end{bmatrix}$$

- Then we plug-and-chug

$$\widehat{se(\hat{\beta})} = \frac{N}{N-k}(X'X)^{-1}X'\,\hat{\Omega}\,X(X'X)^{-1}$$

Time for R

Key takeaways

1. Linear regression approximates the CEF
    a. Coefficients can be interpreted as the change in the slope of the CEF for a unit change in the exposure

2. Analytic solution exists to estimating linear regression coefficients

3. Violations of independent and identically distributed errors assumption can be accounted for in estimating standard errors

30-minute break

Fighting the tyranny of *l'homme moyen*

i.e., a gentle introduction to conditional quantile regression

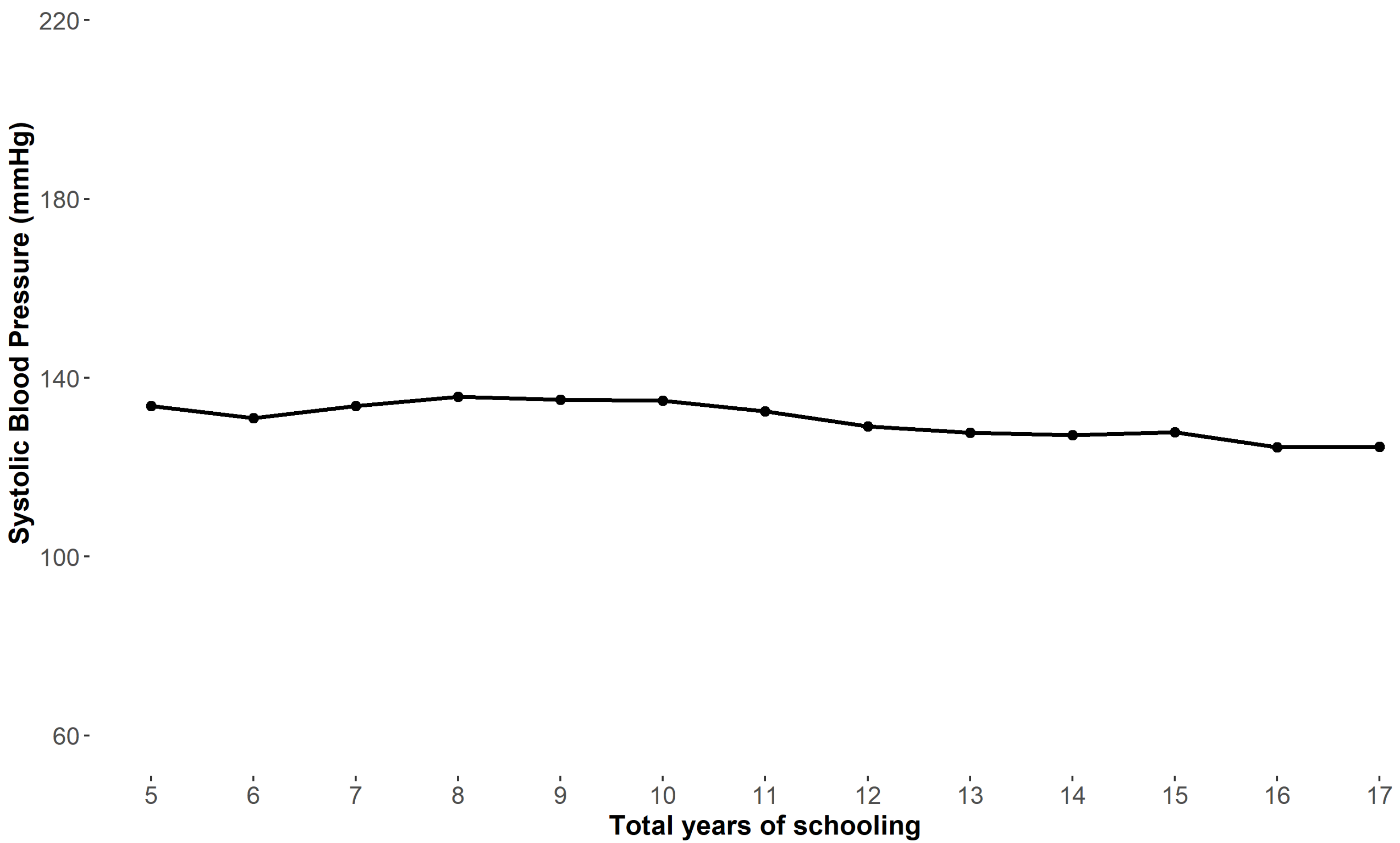
Learning aims

1. Conditional quantile regressions (CQR) and the conditional quantile function

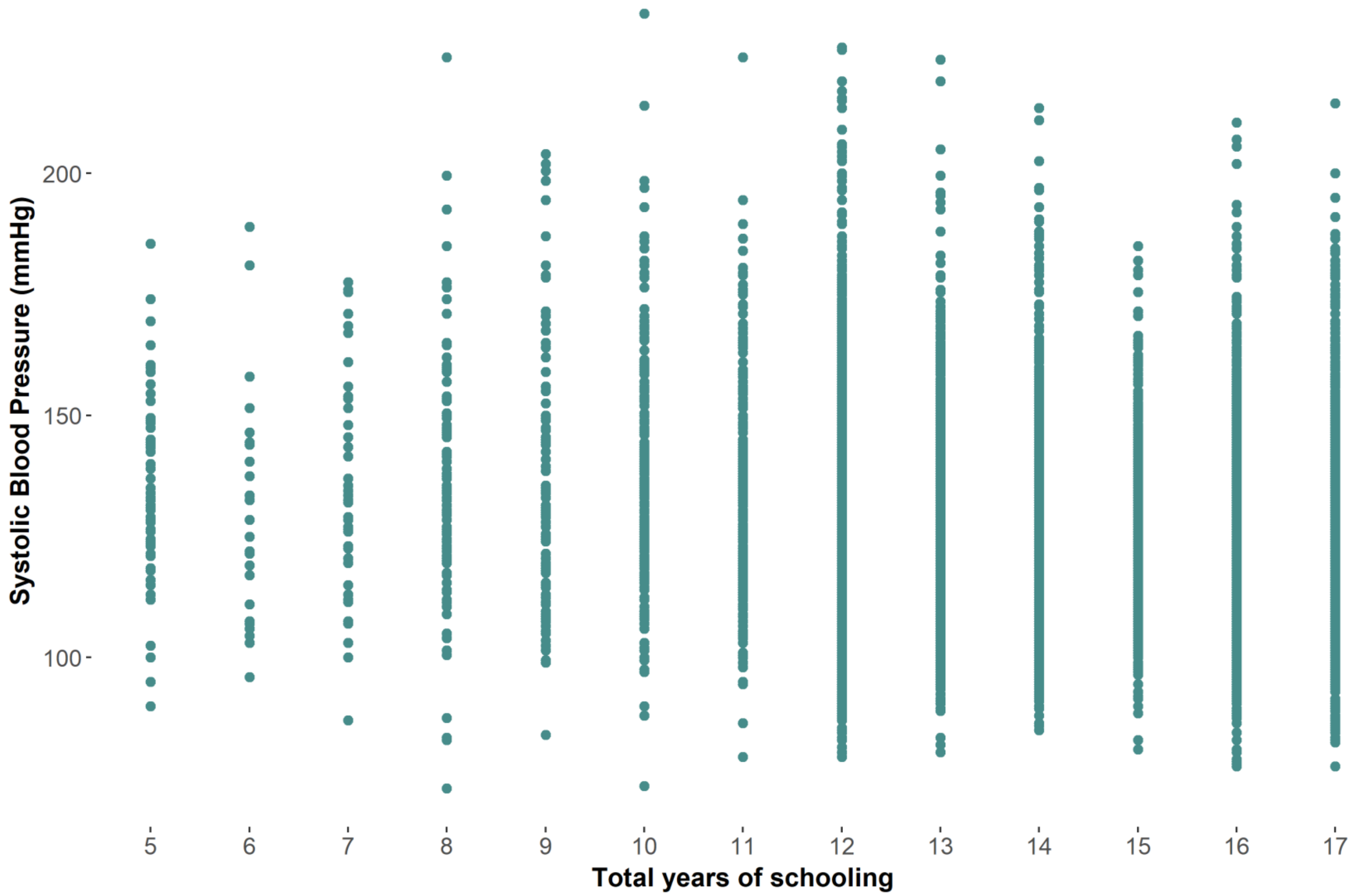
2. Estimating coefficients and standard errors in CQR

3. Interpreting CQR results

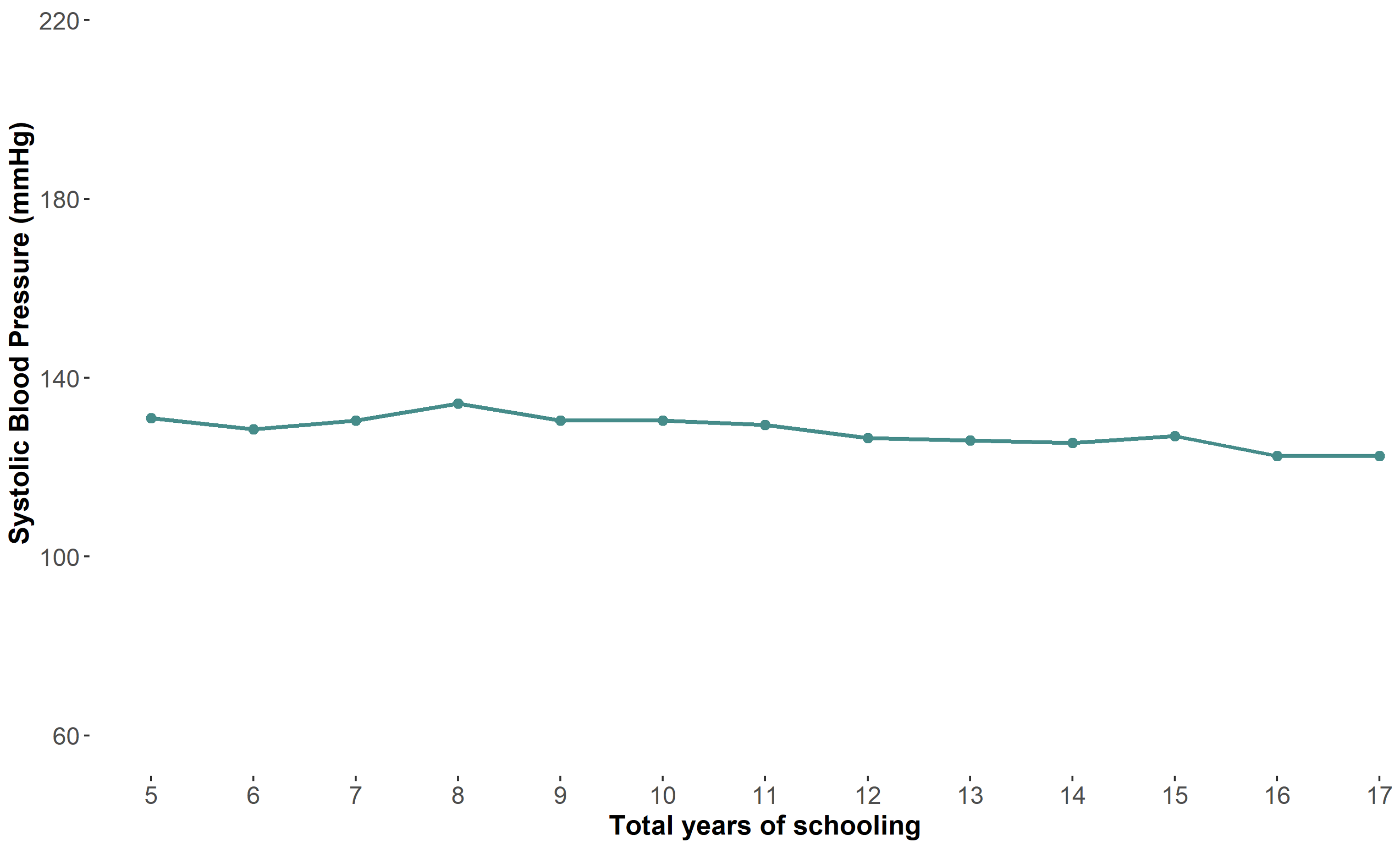
The CEF told us how the conditional mean of SBP changes as schooling increases



Imagine that we calculated the conditional median instead

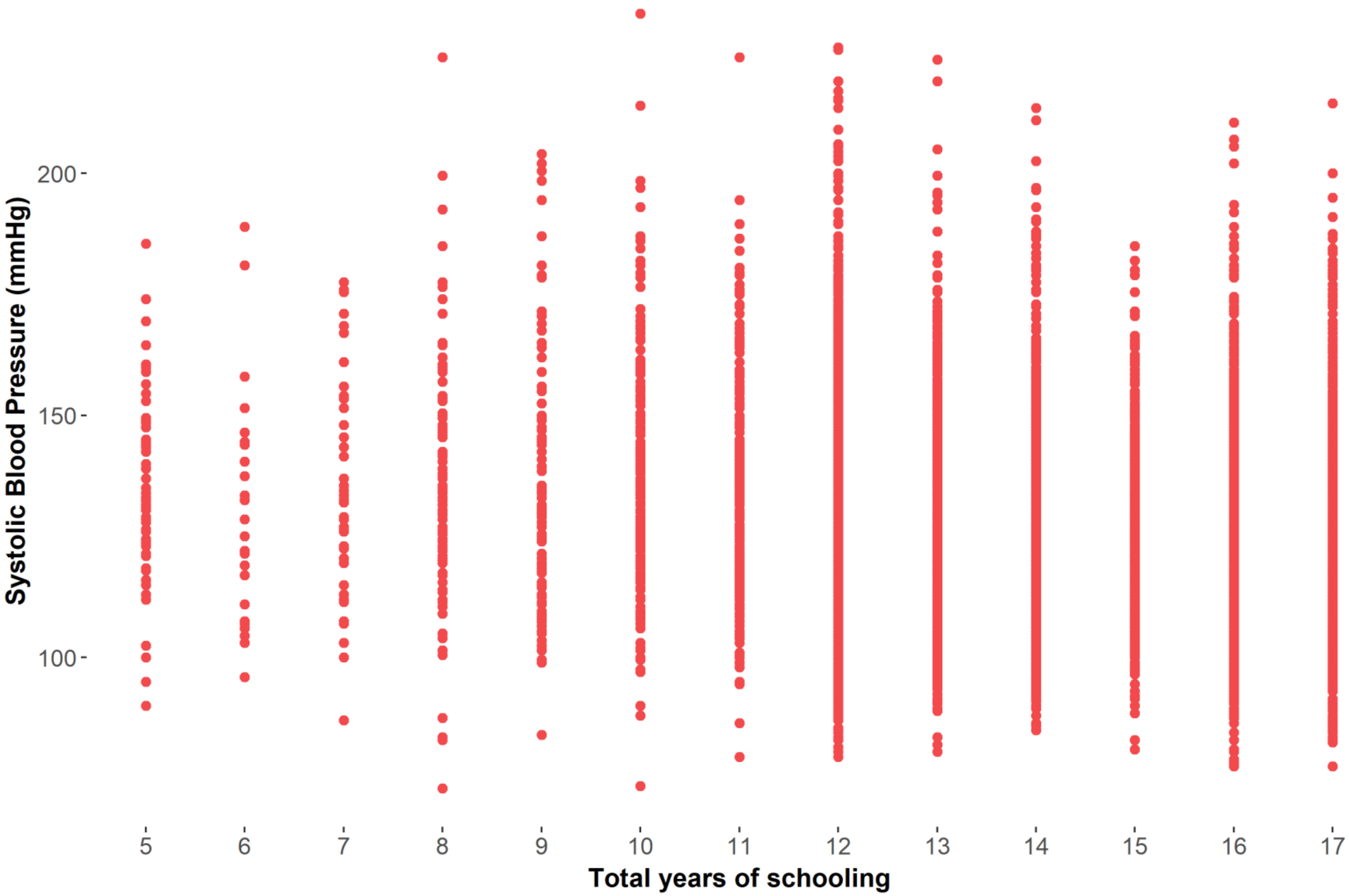


Conditional quantile function (CQF) at the 50th quantile

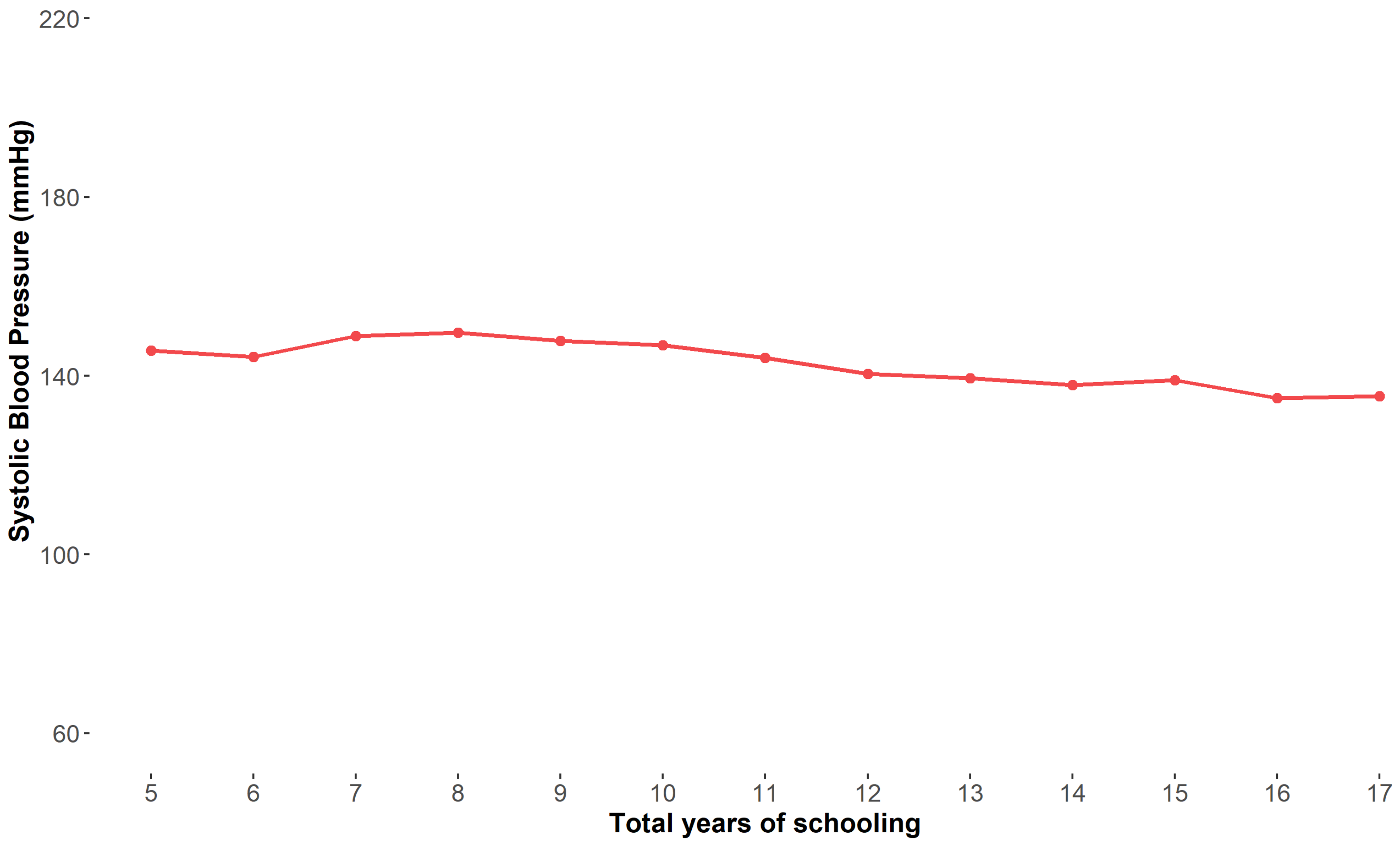


The Conditional Quantile Function at the 50th quantile tells us how the median of the conditional outcome distribution changes as we change values of conditioning variable

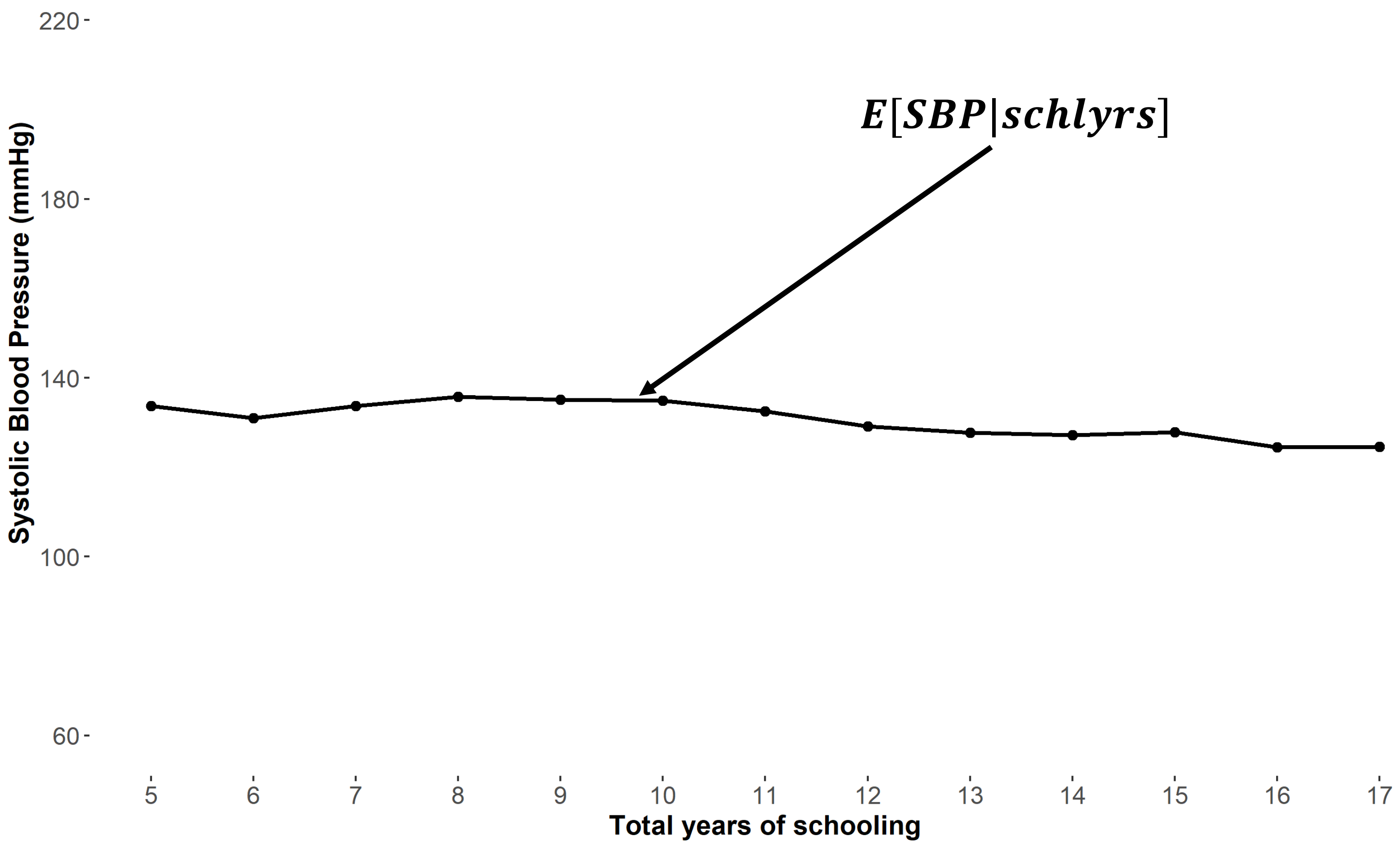
Now imagine that we calculated the conditional 75th quantile



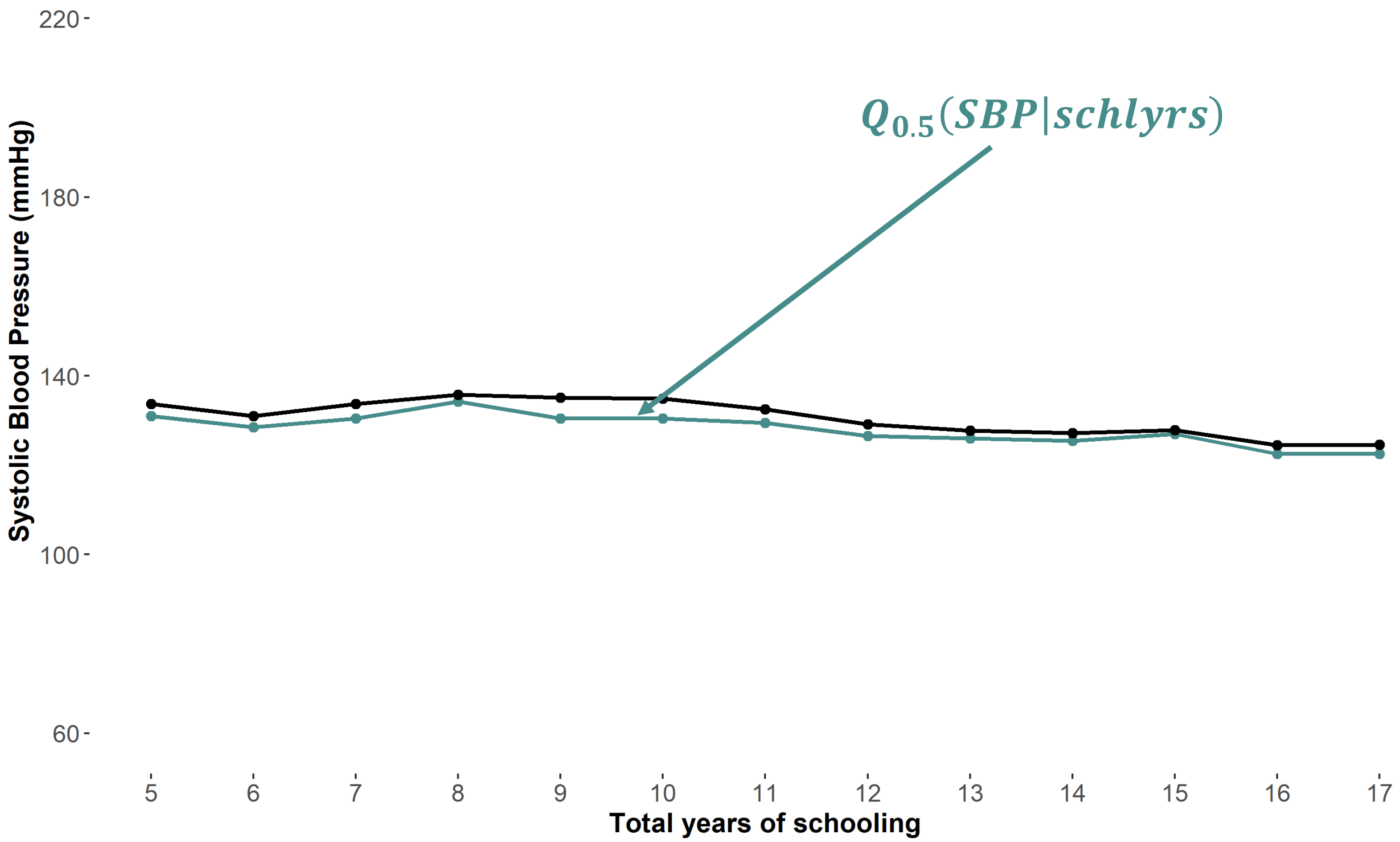
CQF at the 75th quantile



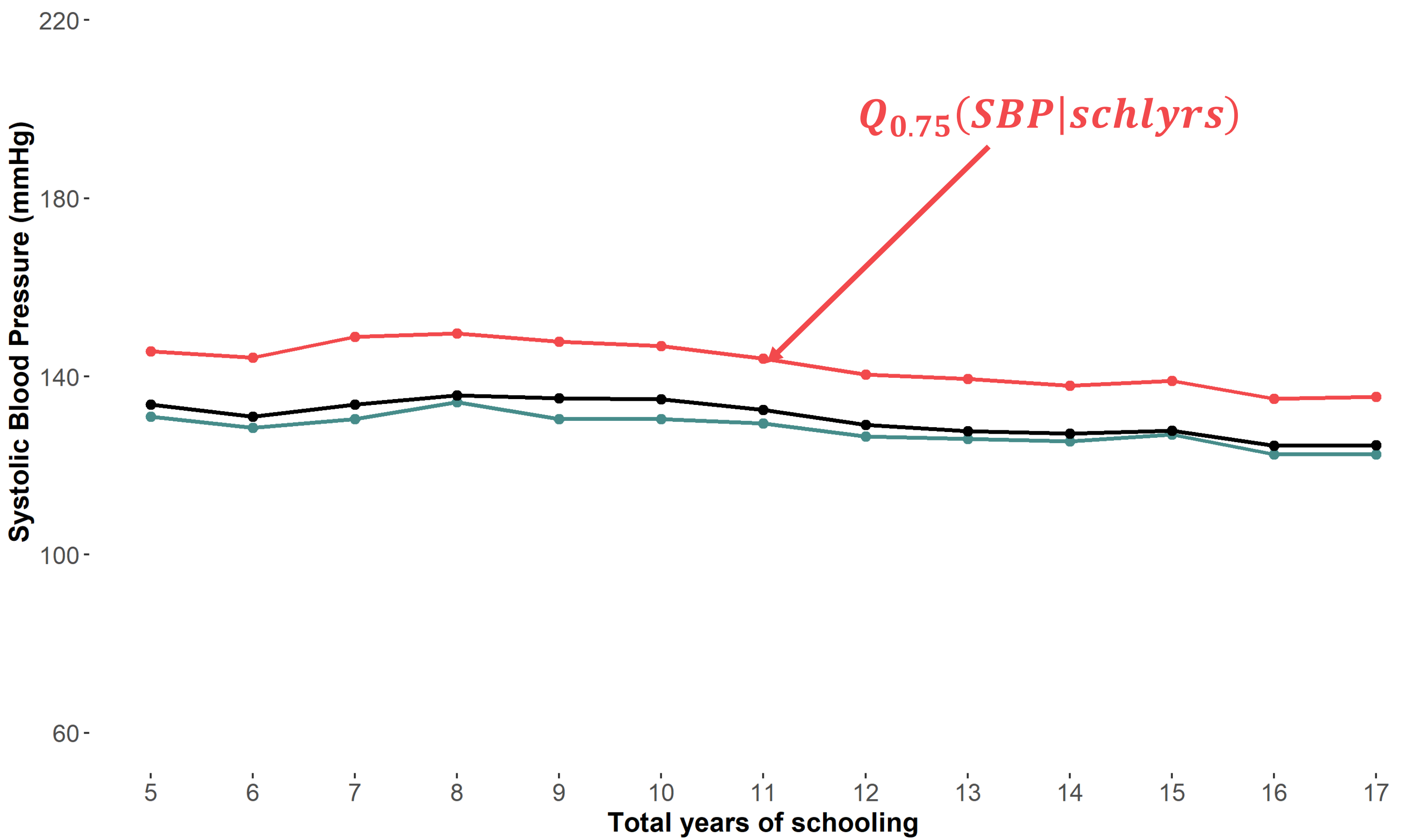
All together!



All together!

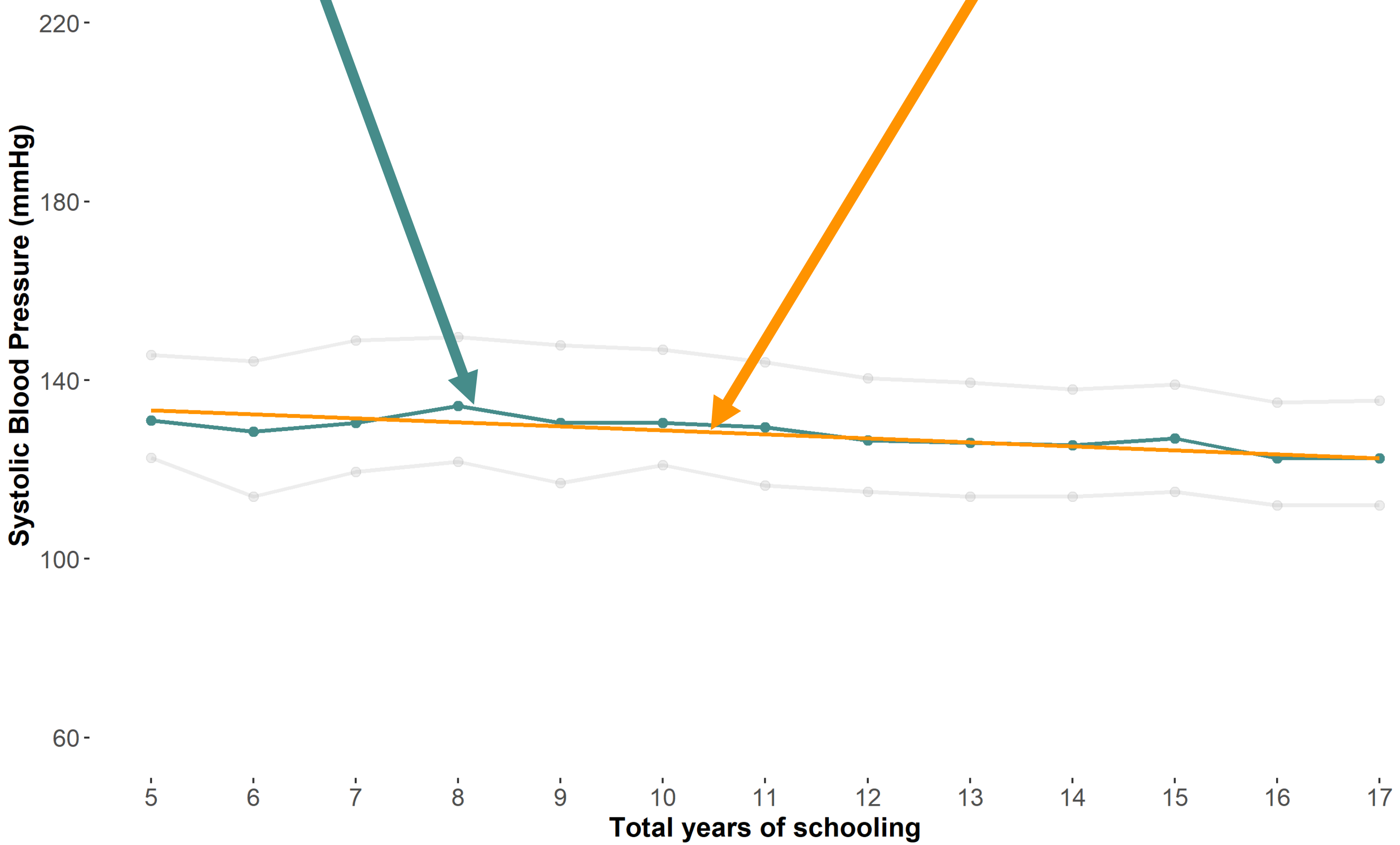


All together!

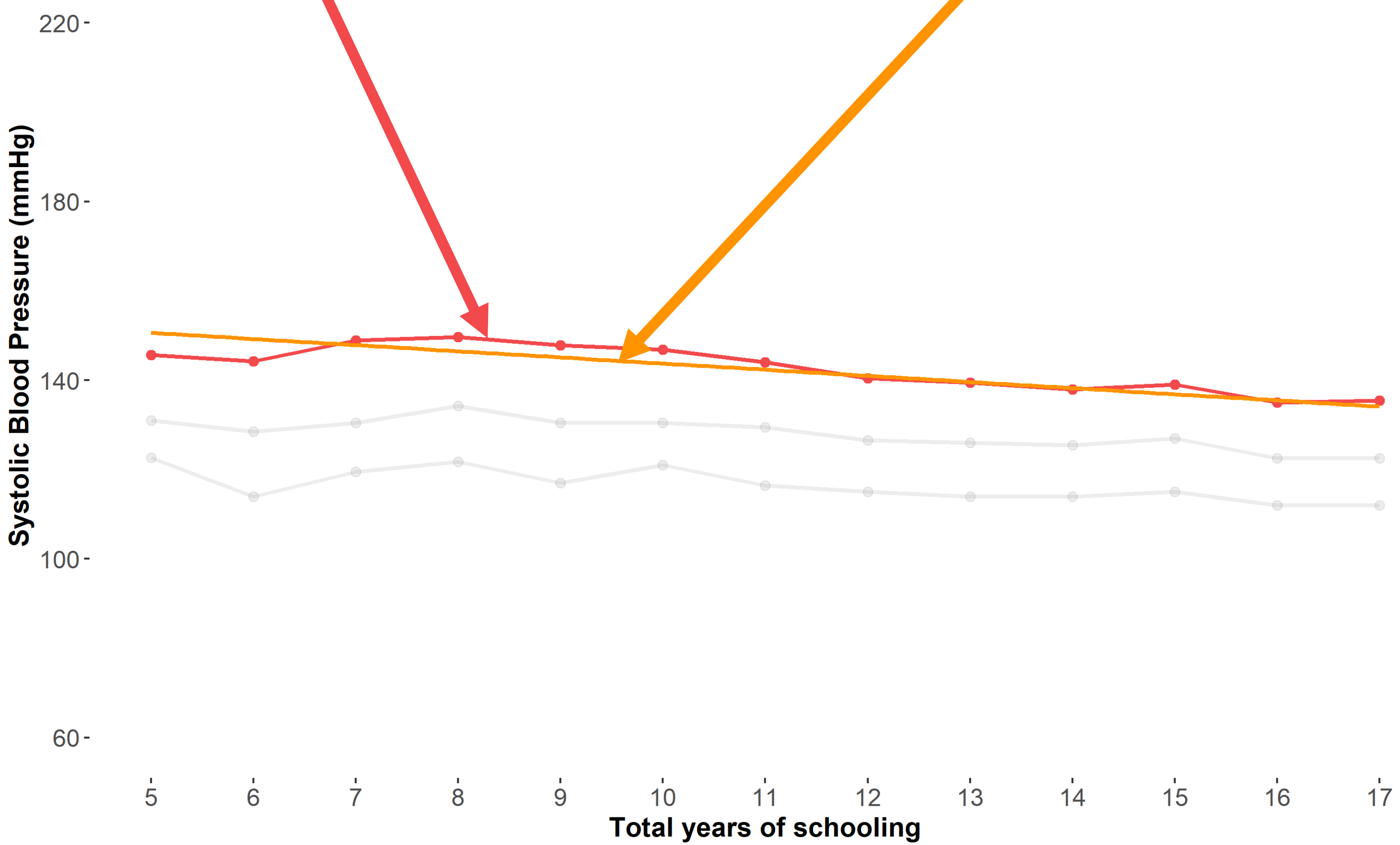


Just as linear regression models the conditional expectation function, conditional quantile regression models the conditional quantile function

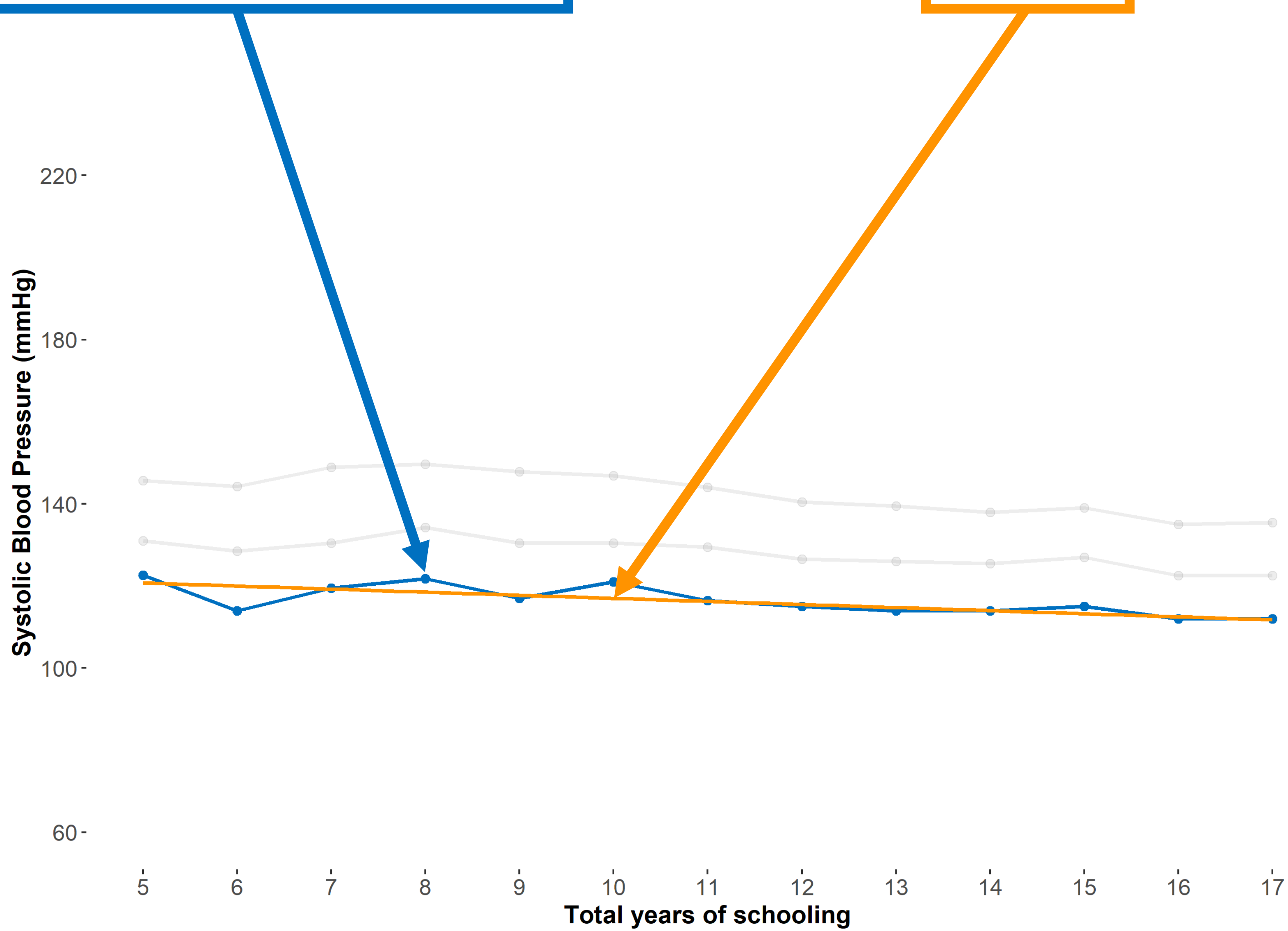
$$\boxed{Q_{0.5}(SBP|schlyrs)} = \alpha_{0,0.5} + \boxed{\alpha_{1,0.5}}schlyrs_i$$



$$\boxed{Q_{0.75}(SBP|schlyrs)} = \alpha_{0,0.75} + \boxed{\alpha_{1,0.75}}\, schlyrs_i$$



$$\boxed{Q_{0.25}(SBP|schlyrs)} = \alpha_{0,0.25} + \boxed{\alpha_{1,0.25}} schlyrs_i$$

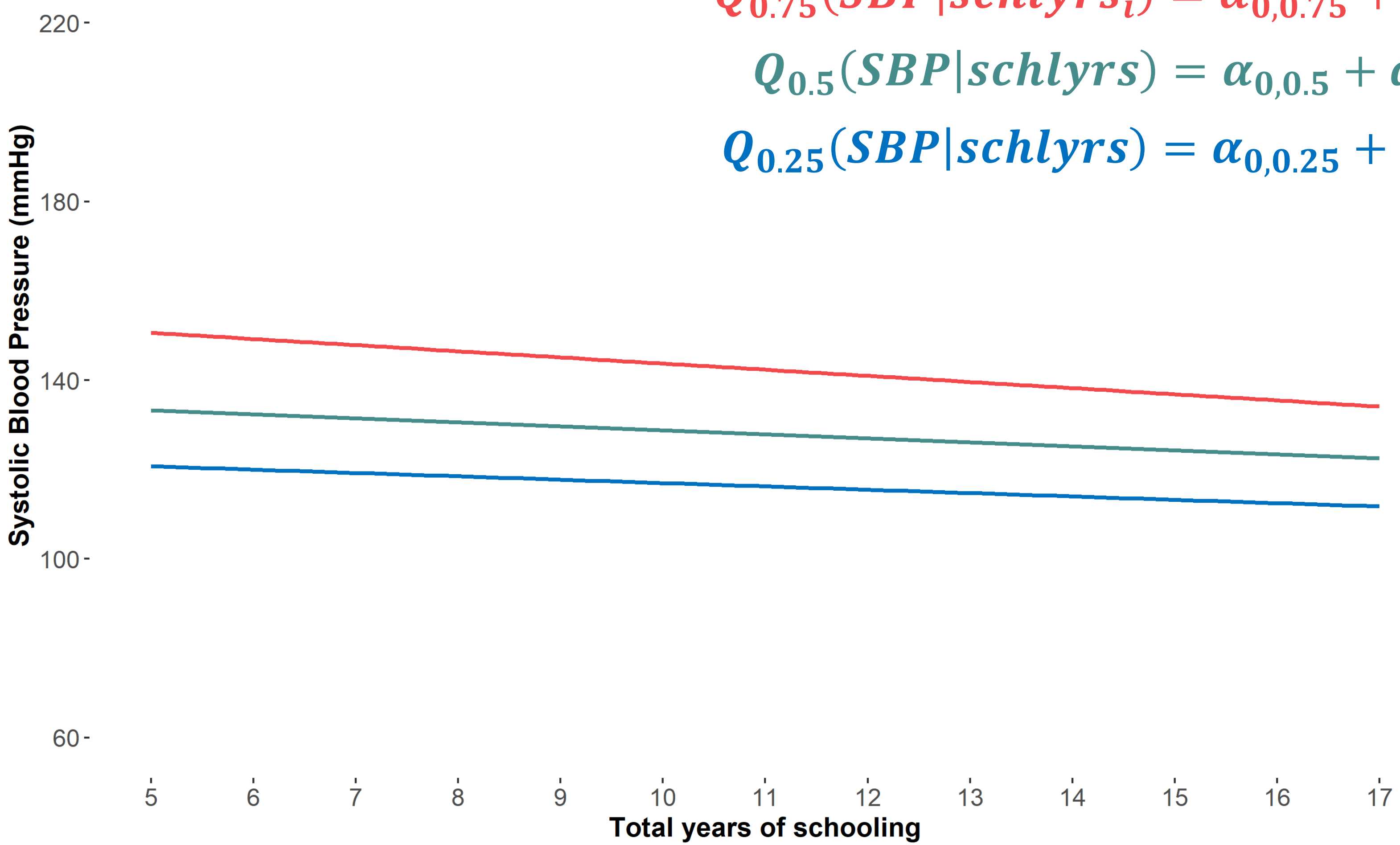


CQR in sample ≈ population CQR line ≈ CQF

$$Q_{0.75}(SBP|schlyrs_i) = \alpha_{0,0.75} + \alpha_{1,0.75}schlyrs_i$$

$$Q_{0.5}(SBP|schlyrs) = \alpha_{0,0.5} + \alpha_{1,0.5}schlyrs_i$$

$$Q_{0.25}(SBP|schlyrs) = \alpha_{0,0.25} + \alpha_{1,0.25}schlyrs_i$$

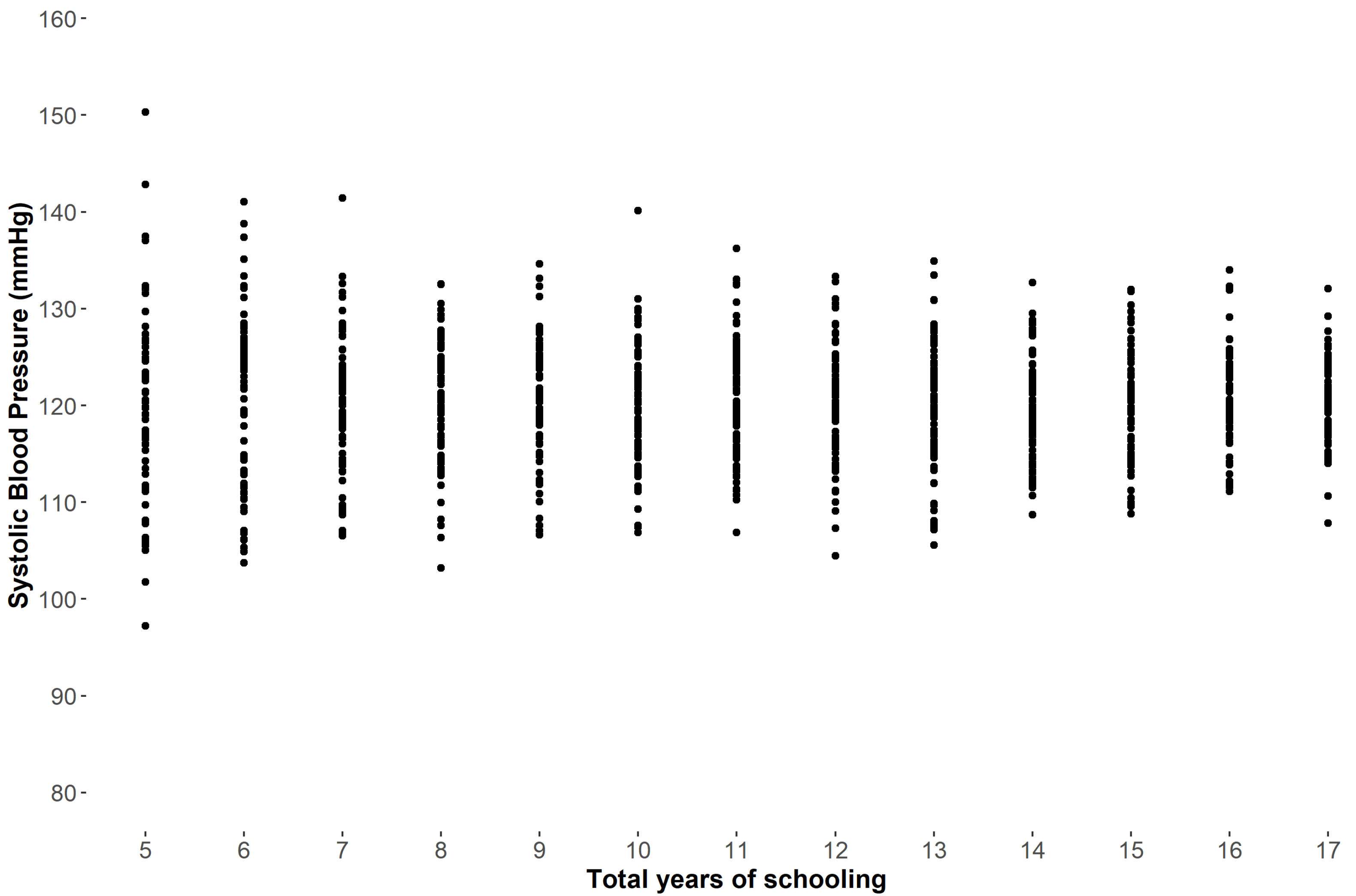


Since it is a model of the conditional quantile function, conditional quantile regression models how the τ^{th} quantile of the outcome changes as we change the exposure by one unit

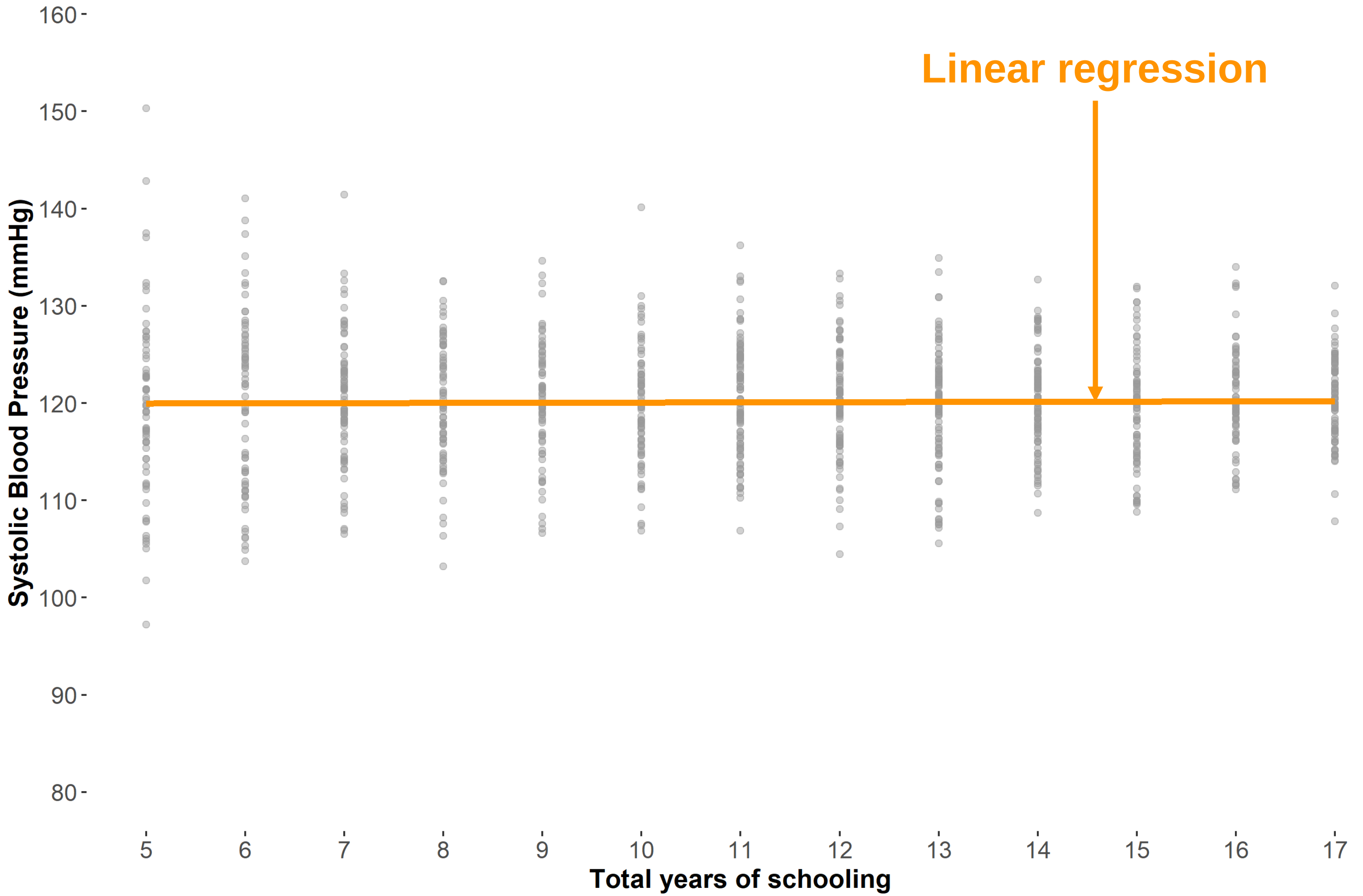
In the context of our example, conditional quantile regression answers the question: By how much does the τ^{th} quantile of SBP change for each additional year of schooling?

Why should we care about conditional quantile regression?

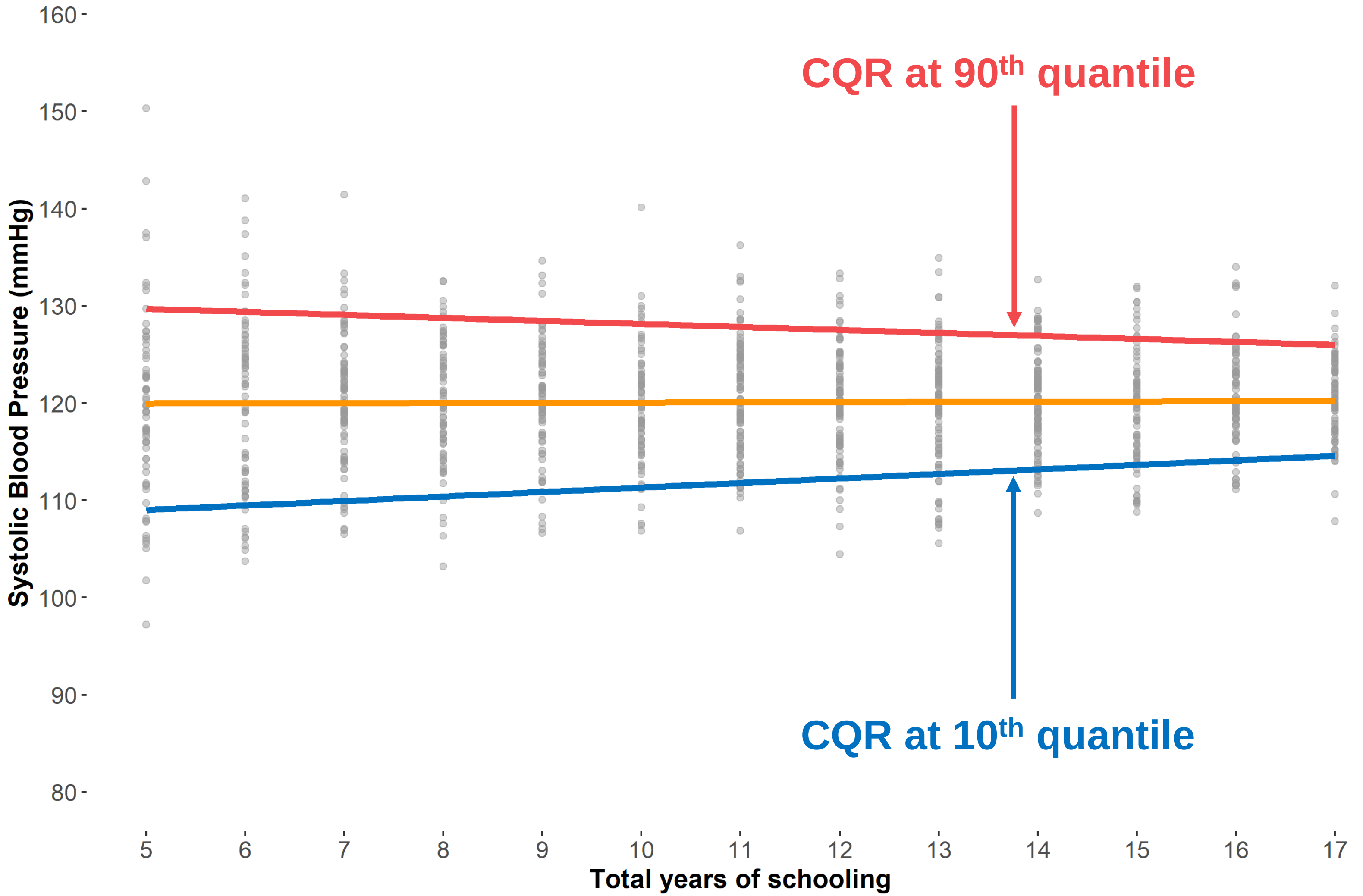
Quantile regressions can quantify distributional effects



Quantile regressions can quantify distributional effects



Quantile regressions can quantify distributional effects



Quantile regressions are robust 💪💪💪

- Quantile regressions are usually robust to:
    1. Outliers in the outcome
    2. Monotonic transformation of the outcome (e.g., natural log)
    3. Top coded or bottom coded outcomes

How do we estimate coefficients of a conditional quantile regression?

Roger Koenker. Source: link

Gilbert Bassett, Jr. Source: link

Econometrica, Vol. 46, No. 1 (January, 1978)

REGRESSION QUANTILES[1]

BY ROGER KOENKER AND GILBERT BASSETT, JR.

A simple minimization problem yielding the ordinary sample quantiles in the location model is shown to generalize naturally to the linear model generating a new class of statistics we term "regression quantiles." The estimator which minimizes the sum of absolute residuals is an important special case. Some equivariance properties and the joint asymptotic distribution of regression quantiles are established. These results permit a natural generalization to the linear model of certain well-known robust estimators of location.

Estimators are suggested, which have comparable efficiency to least squares for Gaussian linear models while substantially out-performing the least-squares estimator over a wide class of non-Gaussian error distributions.

Koenker and Bassett (1978). Regression Quantiles. Econometrica 46(1): 33-50

Population CQR line ≈ CQF

- Recall that the population CQR line is a model of the CQF for the τ^{th} quantile, where $\tau \in (0,1)$

$$Q_\tau(Y|X) = \beta_{0,\tau} + \beta_{1,\tau} x_i$$

- Here, we assume that

$$Q_\tau(\epsilon|X) = 0$$

Coefficients of the population CQR line

- Calculate the coefficients of the population CQR by minimizing the sum of asymmetrically weighted absolute errors (i.e., $\rho_\tau(\epsilon_i)$)

$$\left(\beta_{0,\tau}, \beta_{1,\tau}\right) = \min \sum_i \rho_\tau(\epsilon_i) = \min_{\beta_{0,\tau}, \beta_{1,\tau}} \sum_i \rho_\tau(y_i - \beta_{0,\tau} - \beta_{1,\tau} x_i)$$

- Notice that this is how we find the conditional quantiles!

$$Q_\tau(Y|X) = \min_{\beta_\tau} \sum_i \rho_\tau\left(y_i - g(x_i, \beta_\tau)\right)$$

Estimating CQR coefficients in our sample

- We can use estimates of the error term to estimate CQR coefficients

$$\left(\widehat{\beta_{0,\tau}}, \widehat{\beta_{1,\tau}}\right) = \min \sum_i \rho_\tau(\widehat{\epsilon_i}) = \min_{\widehat{\beta_{0,\tau}}, \widehat{\beta_{1,\tau}}} \sum_i \rho_\tau\left(y_i - \widehat{\beta_{0,\tau}} - \widehat{\beta_{1,\tau}} x_i\right)$$

- Unlike OLS, the CQR minimization problem does not have an analytic solution
  - However, this can be solved using linear programming algorithms!

Interpreting coefficients of a univariate CQR

$$\widehat{Q_\tau(Y|X)} = \widehat{\beta_{0,\tau}} + \widehat{\beta_{1,\tau}} x_i$$

- $\widehat{\beta_{0,\tau}}$ is the estimated τ^{th} quantile of Y when $x_i = 0$ (i.e., the τ^{th} quantile of the distribution $Y|x_i = 0$)

- $\widehat{\beta_{1,\tau}}$ is the estimated change in the τ^{th} quantile of the conditional distribution of Y for a one-unit change in X
  - If $X = \{0,1\}$, then $\widehat{\beta_{1,\tau}}$ is our estimate of the difference in the τ^{th} quantile of Y when $x_i = 1$ and $x_i = 0$

Apply caution when interpreting CQR results!

ID	$Y(x_i = 0)$	$Y(x_i = 1)$
1	13	4
2	14	5
3	15	2
4	2	8
5	5	7
6	11	5
7	17	3
8	6	9
9	12	6
$\overline{Y(X = x)}$	10.6	5.4

Linear regression estimates:

$$\overline{Y(x_i = 1)} - \overline{Y(x_i = 0)}$$

$$= 5.4 - 10.6 = -5.2.$$

Apply caution when interpreting CQR results!

ID	$Y(x_i = 0)$	$Y(x_i = 1)$	$Y(x_i = 1) - Y(x_i = 0)$
1	13	4	-9
2	14	5	-9
3	15	2	-13
4	2	8	6
5	5	7	-2
6	11	5	-6
7	17	3	-14
8	6	9	-3
9	12	6	6
$\overline{Y(X = x)}$	10.6	5.4	

Apply caution when interpreting CQR results!

ID	$Y(x_i = 0)$	$Y(x_i = 1)$	$Y(x_i = 1) - Y(x_i = 0)$
1	13	4	-9
2	14	5	-9
3	15	2	-13
4	2	8	6
5	5	7	-2
6	11	5	-6
7	17	3	-14
8	6	9	-3
9	12	6	6
$\overline{Y(x_i = 1) - Y(x_i = 0)}$	10.6	5.4	**-5.2**

Linear regression estimates:

$\overline{Y(x_i = 1)} - \overline{Y(x_i = 0)}$

$= 5.4 - 10.6 = \mathbf{-5.2}$.

Linear regression estimate can be interpreted as average change in Y as an **individual** goes from $x_i = 0$ to $x_i = 1$

Apply caution when interpreting CQR results!

ID	$Y(x_i = 0)$	$Y(x_i = 1)$
1	13	4
2	14	5
3	15	2
4	2	8
5	5	7
6	11	5
7	17	3
8	6	9
9	12	6
$Q_{0.5}\big(Y(X = x)\big)$	12	5

Quantile regression at the median estimates:

$$Q_{0.5}\big(Y(x_i = 1)\big) - Q_{0.5}\big(Y(x_i = 0)\big)$$

$$= 5 - 12 = -7.$$

Apply caution when interpreting CQR results!

ID	$Y(x_i = 0)$	$Y(x_i = 1)$
1	13	4
2	14	5
3	15	2
4	2	8
5	5	7
6	11	5
7	17	3
8	6	9
$\boxed{9}$	$\boxed{12}$	$\boxed{6}$
$Q_{0.5}\big(Y(X = x)\big)$	12	5

Quantile regression at the median estimates:

$$Q_{0.5}\big(Y(x_i = 1)\big) - Q_{0.5}\big(Y(x_i = 0)\big)$$

$$= 5 - 12 = -7.$$

Change in Y as we go from $X = 0$ to $X = 1$ for the individual at the median under $Y(x_i = 0)$:

$$= 6 - 12 = \mathbf{-6}.$$

Apply caution when interpreting CQR results!

ID	$Y(x_i = 0)$	$Y(x_i = 1)$
1	13	4
2	14	5
3	15	2
4	2	8
5	5	7
6	11	5
7	17	3
8	6	9
9	12	6
$Q_{0.5}(Y(X = x))$	12	5

Quantile regression at the median estimates:

$$Q_{0.5}(Y(x_i = 1)) - Q_{0.5}(Y(x_i = 0))$$

$$= 5 - 12 = -7.$$

Change in Y as we go from $X = 0$ to $X = 1$ for the individual at the median under $Y(x_i = 0)$:

$$= 6 - 12 = -6.$$

Change in Y as we go from $X = 0$ to $X = 1$ for the individual at the median under $Y(x_i = 1)$:

$$= 5 - 11 = -6.$$

Conditional quantile regression estimates cannot usually be interpreted as the estimated change in the outcome for an individual at the τ^{th} quantile of the distribution under a specific treatment value

How do we estimate standard errors in conditional quantile regression?

A heteroskedastic world

- When the errors are dependent on the covariates, we have a "robust" variance formula

$$V\left(\hat{\beta}_\tau\right) = \tau(1-\tau)Q_\tau^{-1} Q Q_\tau^{-1}$$

- Here,
  - $\tau(1-\tau)$ is a constant for any given τ
  - $Q = E[X'X]$
  - $Q_\tau = E[X'X]f_\tau(0)$

A heteroskedastic world

- Notice that in the variance formula we are taking the inverse of Q_τ

$$V\left(\widehat{\beta_\tau}\right) = \tau(1-\tau)Q_\tau^{-1}QQ_\tau^{-1}$$

- Since $Q_\tau = E[X'X]f_\tau(0)$, where $f_\tau(0)$ is low, standard errors are going to be large
  - In places where the outcome is sparse (usually the tails), standard errors are going to be large

Key criticism of conditional quantile regression is that statistical power is usually a problem at the tails of the outcome distribution, which is substantively of interest

Bootstrapped standard errors

- Despite having these various formulas, recommended practice is to bootstrap

- In general, bootstrap relies on random resampling with replacement of the observed data and fitting CQR in each sample
  - Different bootstrap methods available in the R package

- Standard non-parametric bootstrap is an estimator for the “robust” CQR standard errors

Time for R

Key takeaways

1. Just as linear regression models the CEF, CQR models the CQF
   - The CQF can be defined for any quantile of interest

2. CQR coefficients can be estimated by minimizing the sum of $\rho_\tau(.)$, which does not have an analytic solution
   - Coefficients can be interpreted as the change in the conditional quantile of interest for a unit change in the exposure

3. CQR standard errors are inversely proportional to the error density, because of which in parts of the outcome distribution where outcome data are sparse, standard errors are larger
   - Bootstrap is the preferred method of estimating the standard errors

10-minute break

Fighting the tyranny of *l'homme moyen*: The sequel

i.e., a gentle introduction to unconditional quantile regressions

Learning aims

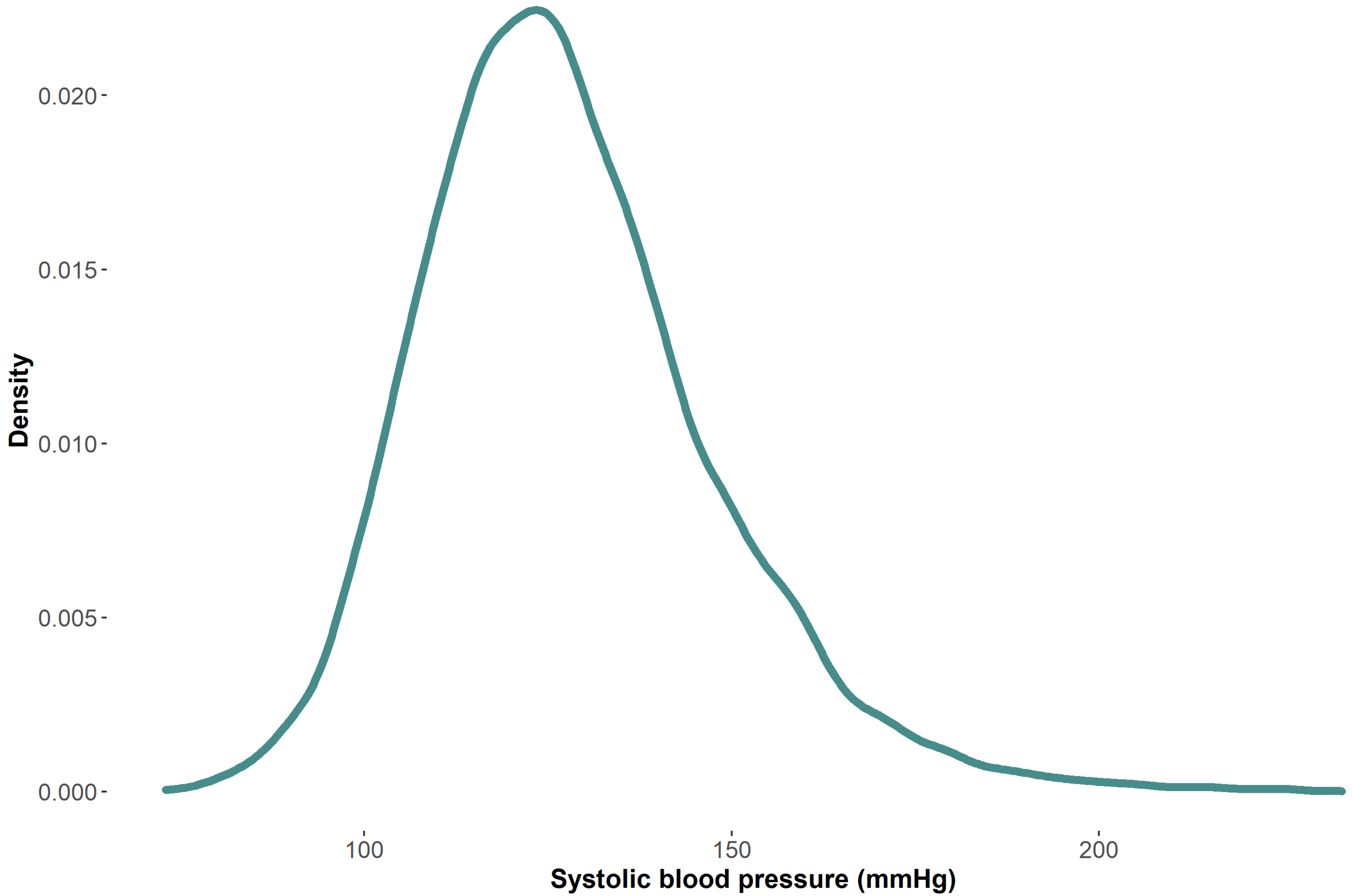
1. Counterfactual unconditional distributions and contrasts

2. Firpo, Fortin, and Lemieux (2009) method of estimating parameters of the unconditional quantile regression

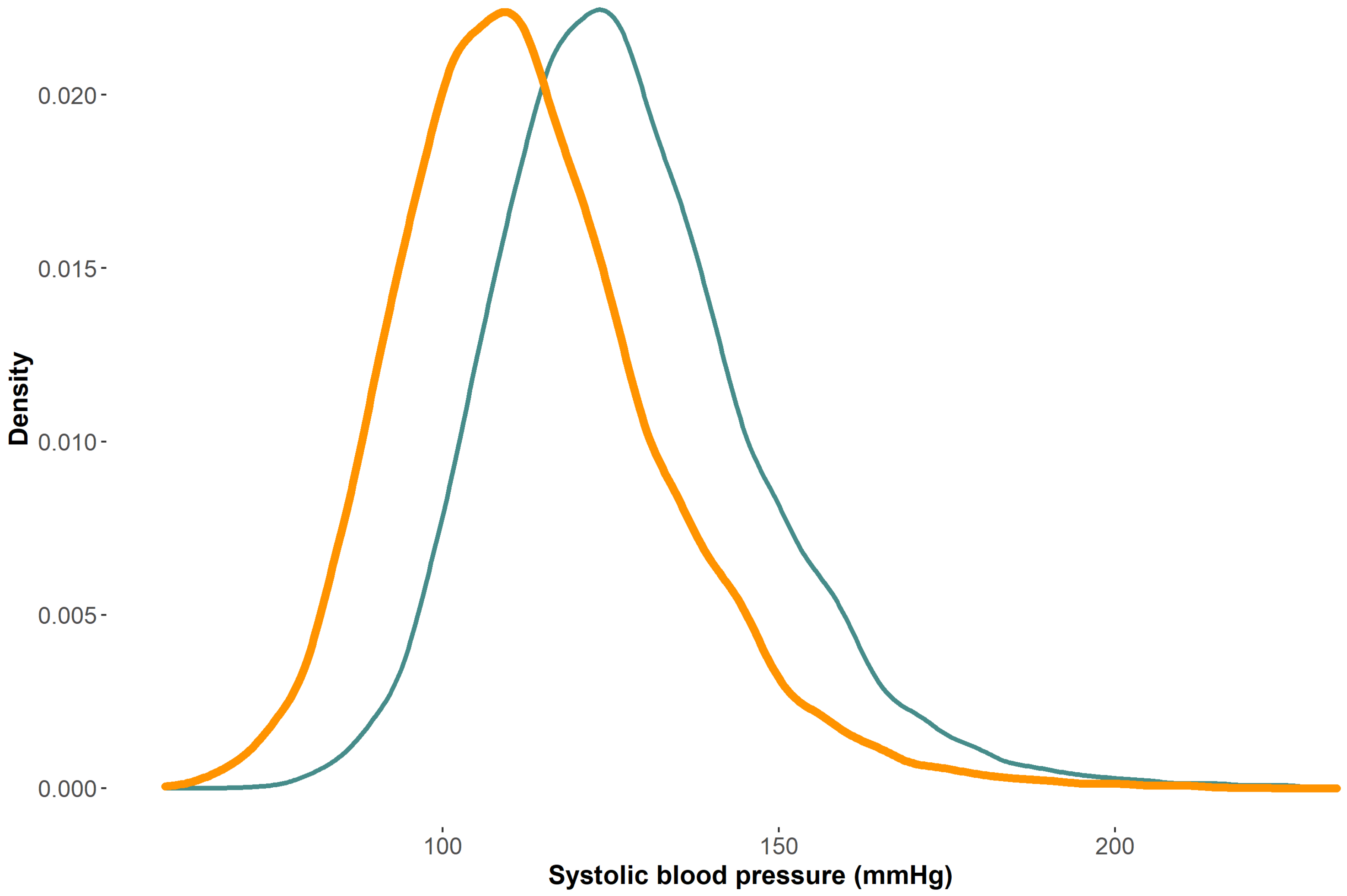
3. Interpreting unconditional quantile regression results

What if we wanted to learn how the unconditional outcome distribution changed in response to some population-level intervention?

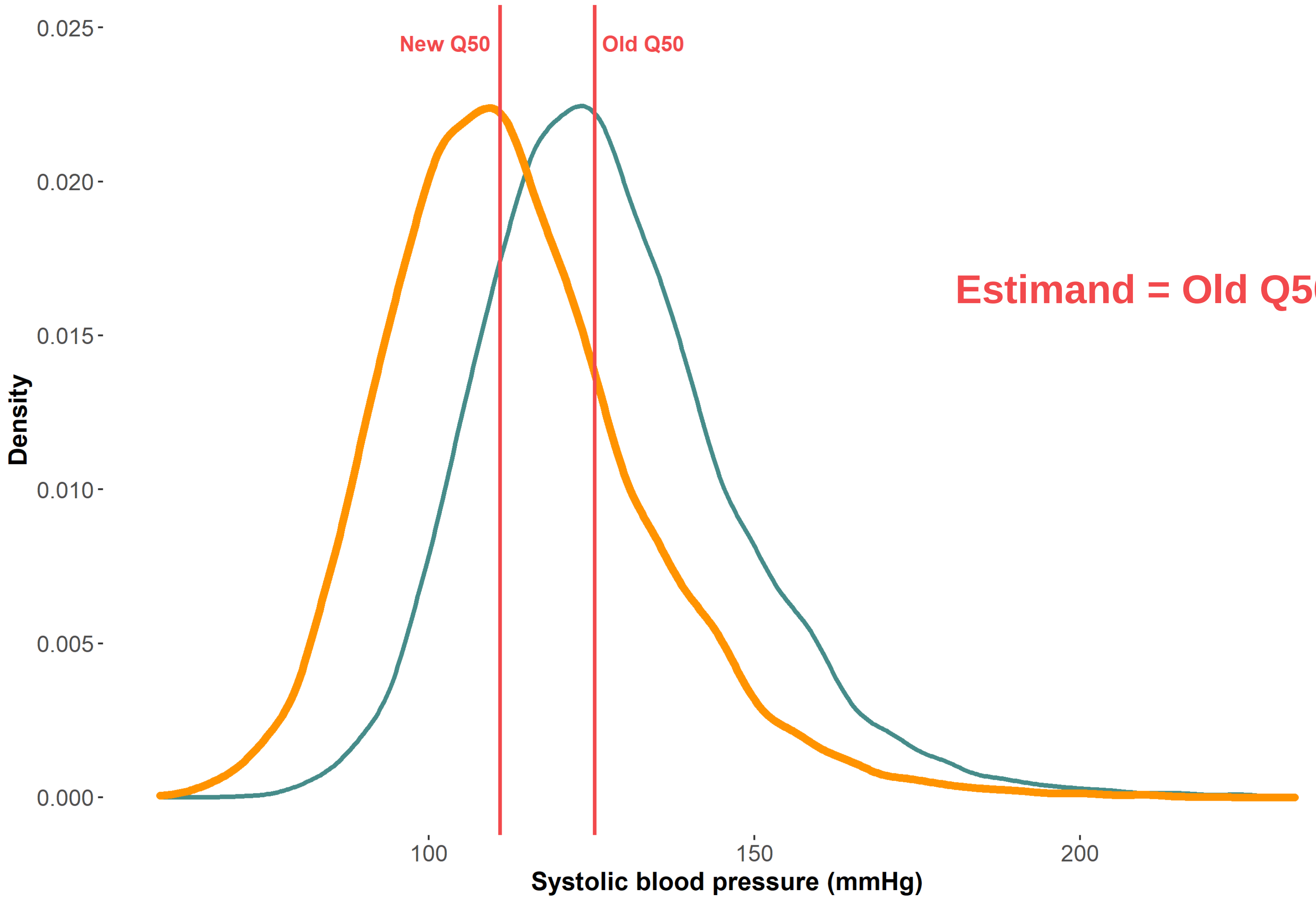
Here's the unconditional distribution of SBP in our data



Suppose everyone gets an intervention creating a counterfactual SBP distribution



We are interested in the difference in quantiles of these distributions



Estimand = Old Q50 – New Q50

Can we use conditional quantile regressions to estimate contrasts of the unconditional outcome distribution?

Can we use conditional quantile regressions to estimate contrasts of the unconditional outcome distribution?

It depends!

According to Borah and Basu (2013)

- Conditional quantile regressions quantify contrasts of the unconditional outcome distribution if:

  1. Outcome is only a function of the exposure
  2. Exposure induces only a location shift in the outcome in the presence of other covariates

- But not if

  1. Exposure interacts with other covariates in the DGP

Is all hope lost?

Sergio Firpo. Source: link

Nicole Fortin. Source: link

Thomas Lemieux. Source: link

Econometrica, Vol. 77, No. 3 (May, 2009), 953–973

UNCONDITIONAL QUANTILE REGRESSIONS

BY SERGIO FIRPO, NICOLE M. FORTIN, AND THOMAS LEMIEUX[1]

We propose a new regression method to evaluate the impact of changes in the distribution of the explanatory variables on quantiles of the unconditional (marginal) distribution of an outcome variable. The proposed method consists of running a regression of the (recentered) influence function (RIF) of the unconditional quantile on the explanatory variables. The influence function, a widely used tool in robust estimation, is easily computed for quantiles, as well as for other distributional statistics. Our approach, thus, can be readily generalized to other distributional statistics.

Firpo, Fortin, Lemieux (2009). Unconditional Quantile Regressions. Econometrica 77(3): 953-973

Firpo's unconditional quantile regression method quantifies the change in the τ^{th} quantile of the unconditional outcome distribution for a small change in the exposure distribution

How does Firpo et. al (2009) quantify the change in the τ^{th} quantile of the unconditional outcome distribution?

Recentered Influence Function (RIF) to the rescue!

- Recentered influence function (RIF): sum of a distributional statistic's influence function and the distributional statistic itself

- RIF regression: regression of the RIF on the exposure and other covariates of interest
  - Effect on quantile of the unconditional outcome distribution for a small shift in the unconditional exposure distribution (i.e., the Unconditional Quantile Partial Effect)
  - Several methods of estimating RIF regression including OLS

What’s an Influence Function?

- Influence functions (IF) are a way of assessing the “robustness” of distributional statistics

- IF calculates change in a distributional statistic of interest for a “small” change in the distribution
  - IF calculates the expected change even when we don’t observe the small change
  - Examples of distributional statistics include means, quantiles, Gini coefficients, etc.
  - Each distributional statistic has an influence function specific to it

Influence Function for quantiles

- Influence function for the τ^{th} quantile of Y

$$IF(y; q_\tau) = \frac{\tau - \boldsymbol{I}(y \leq q_\tau)}{f_y(q_\tau)}$$

Where q_τ is the value of Y at the τ^{th} quantile, $I(.)$ is the indicator function, and $f_y(q_\tau)$ is the density of Y at the τ^{th} quantile

What's a RIF?

- Firpo et. al. (2009) define the RIF for the τ^{th} quantile as

$$RIF(y; q_\tau) = q_\tau + IF(y; q_\tau) = q_\tau + \frac{\tau - \boldsymbol{I}(y \leq q_\tau)}{f_y(q_\tau)}$$

- Expected value of the RIF equals the τ^{th} quantile

$$E[RIF(y; q_\tau)] = E\left[q_\tau + \frac{\tau - \boldsymbol{I}(y \leq q_\tau)}{f_y(q_\tau)}\right] = E[q_\tau] + E\left[\frac{\tau - \boldsymbol{I}(y \leq q_\tau)}{f_y(q_\tau)}\right] = q_\tau$$

How does the Firpo et. al. (2009) method work?

Step 1

Estimate the RIF in the unconditional outcome distribution

Step 2

Regress RIF on exposure + covariates

Step 3

Estimate standard errors using bootstrap or sandwich estimator

Step 1

Estimate the RIF in the unconditional outcome distribution

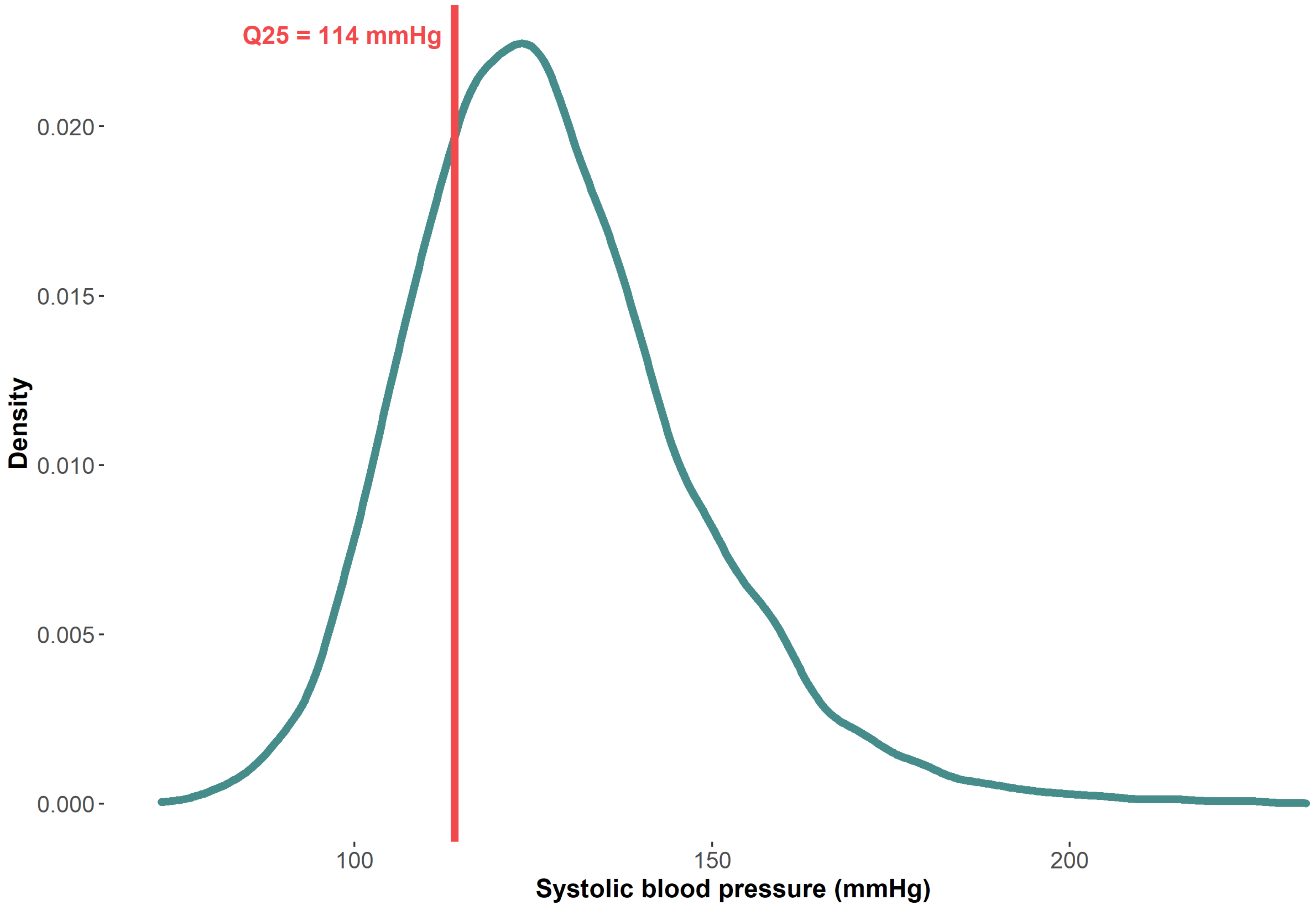
Step 2

Regress RIF on exposure + covariates

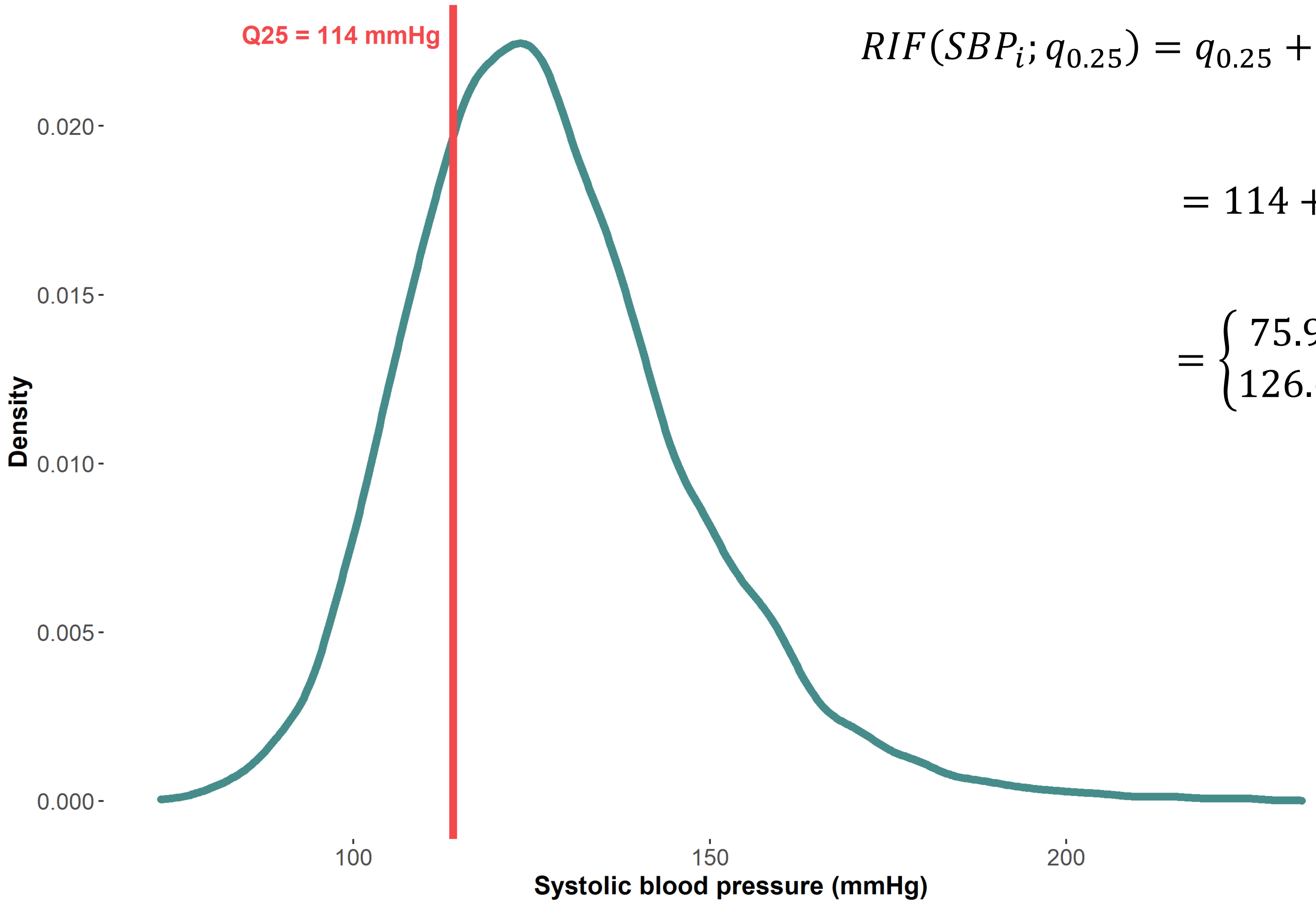
Step 3

Estimate standard errors using bootstrap or sandwich estimator

Step 1: Choose the quantile of interest in the unconditional outcome distribution



Step 1: Estimate the RIF at the quantile of interest

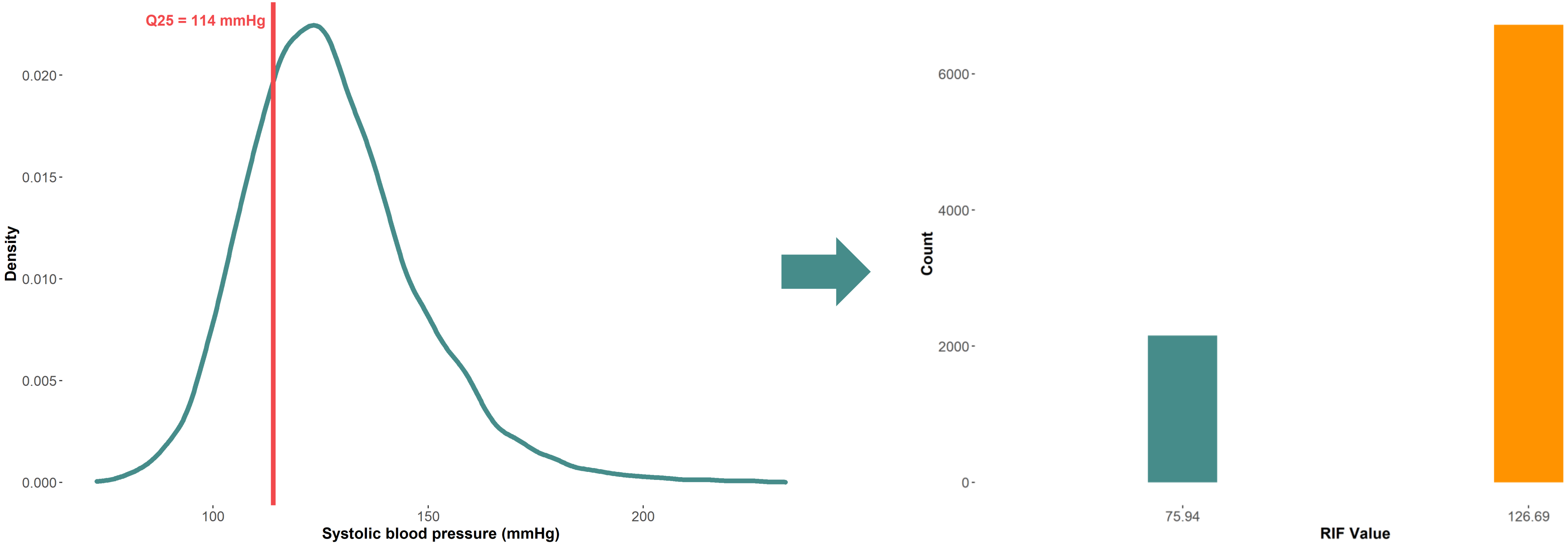


$$RIF(SBP_i; q_{0.25}) = q_{0.25} + \frac{\tau - I(SBP_i \leq q_{0.25})}{f_Y(q_{0.25})}$$

$$= 114 + \frac{0.25 - I(SBP_i \leq 114)}{f_Y(114)}$$

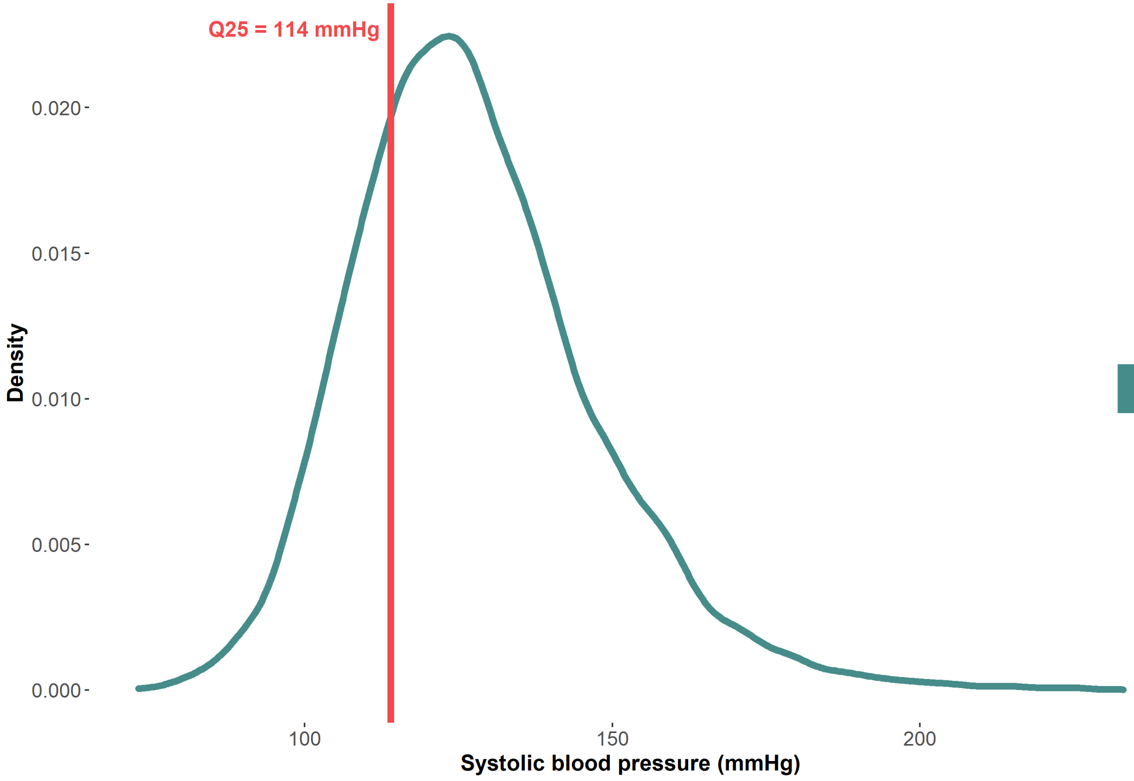
$$= \begin{cases} 75.94; & I(SBP_i \leq 114) \\ 126.69; & I(SBP_i > 114) \end{cases}$$

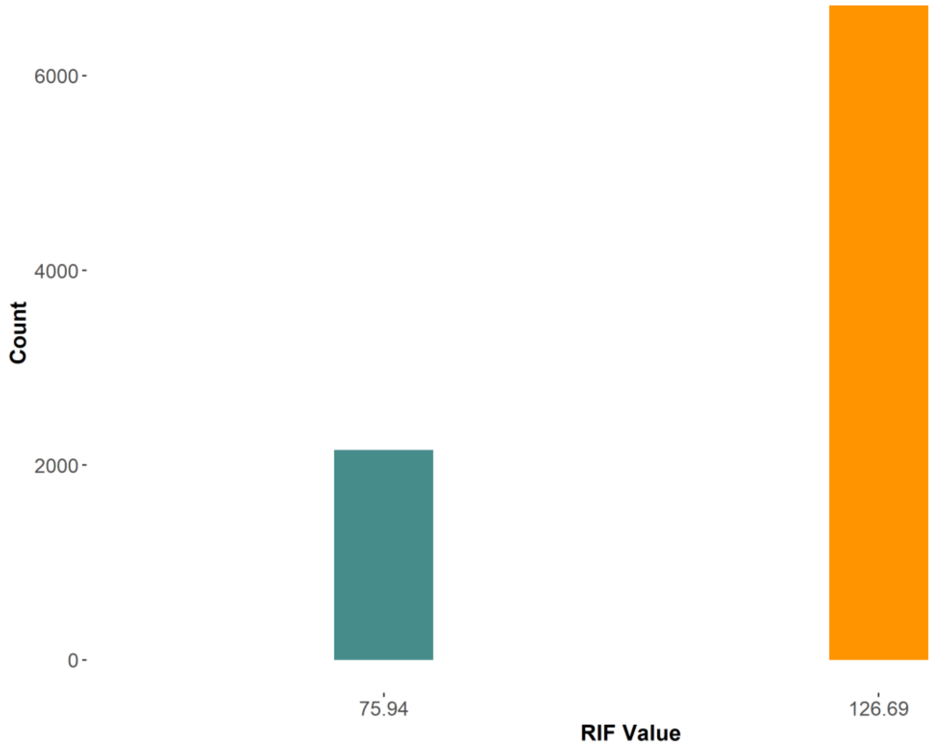
Moving from the unconditional distribution to the RIF



Step 1: Estimated RIF at the 25th quantile of SBP

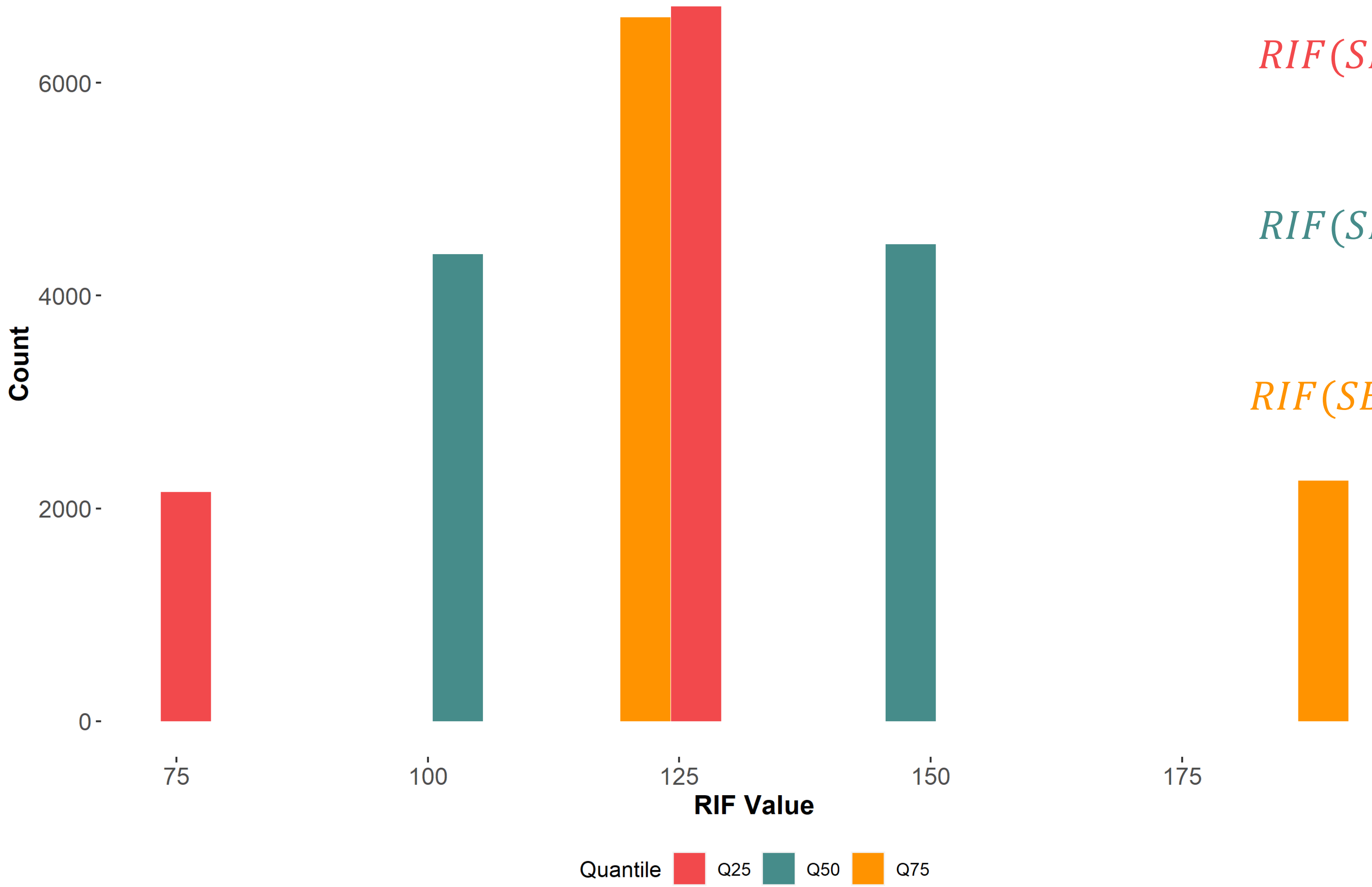
$$RIF(SBP_i, q_{0.25}) = \begin{cases} 75.94; \ I(SBP \leq 114) \\ 126.69; \ I(SBP > 114) \end{cases}$$





$$E[RIF(SBP_i, q_{0.25})] = (0.25)(75.94) + (0.75)(126.69) = 114 \, mmHg = q_{0.25}$$

Step 1: Estimated RIF at the 25th, 50th, and 75th quantiles of SBP



$$RIF(SBP_i, q_{0.25}) = \begin{cases} 75.94;\ I(SBP \leq 114) \\ 126.69;\ I(SBP > 114) \end{cases}$$

$$RIF(SBP_i, q_{0.5}) = \begin{cases} 103.00;\ I(SBP \leq 125.5) \\ 148.00;\ I(SBP > 125.5) \end{cases}$$

$$RIF(SBP_i, q_{0.75}) = \begin{cases} 121.65;\ I(SBP \leq 138.5) \\ 189.05;\ I(SBP > 138.5) \end{cases}$$

Step 1

Estimate the RIF in the unconditional outcome distribution

Step 2

Regress RIF on exposure + covariates

Step 3

Estimate standard errors using bootstrap or sandwich estimator

Step 2: Three ways to regress RIF on exposure and covariates

$$RIF(y;q_\tau) = q_\tau + \frac{\tau - \boldsymbol{I}(y \le q_\tau)}{f_y(q_\tau)} = \frac{1}{f_y(q_\tau)}\boldsymbol{I}(y > q_\tau) + q_\tau - \frac{1-\tau}{f_y(q_\tau)}$$

- Three ways of regressing RIF on exposure and covariates:
  1. Linear regression using OLS with RIF as the outcome (RIF-OLS)
  2. Logistic regression using $\boldsymbol{I}(y > q_\tau)$ as the outcome (RIF-Logit)
  3. Polynomial regression to model $\boldsymbol{I}(y > q_\tau)$ (RIF-NP)

- We will focus on RIF-OLS as it is easiest and most practical

Step 2: Fitting the RIF-OLS regression for SBP at the 25th quantile

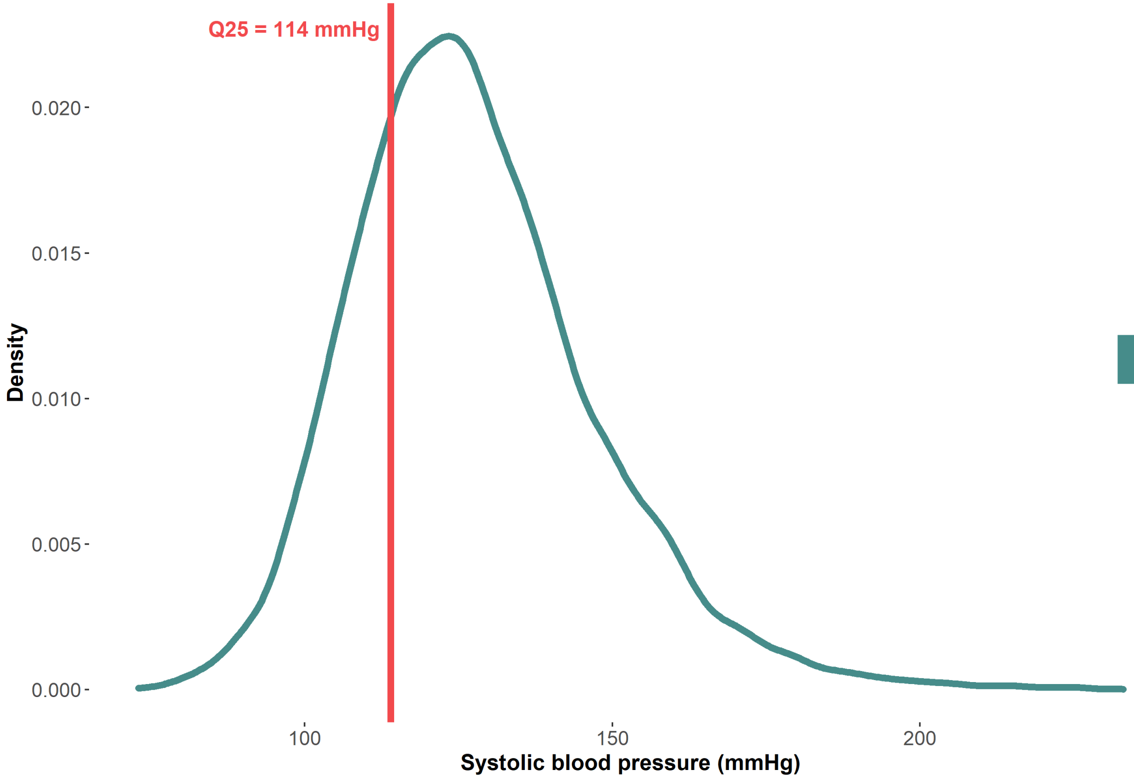
$$E[RIF(SBP_i; q_{0.25})|X, C] = \alpha_{0,0.25} + \alpha_{1,0.25}x_i + \lambda'_{0.25}C$$

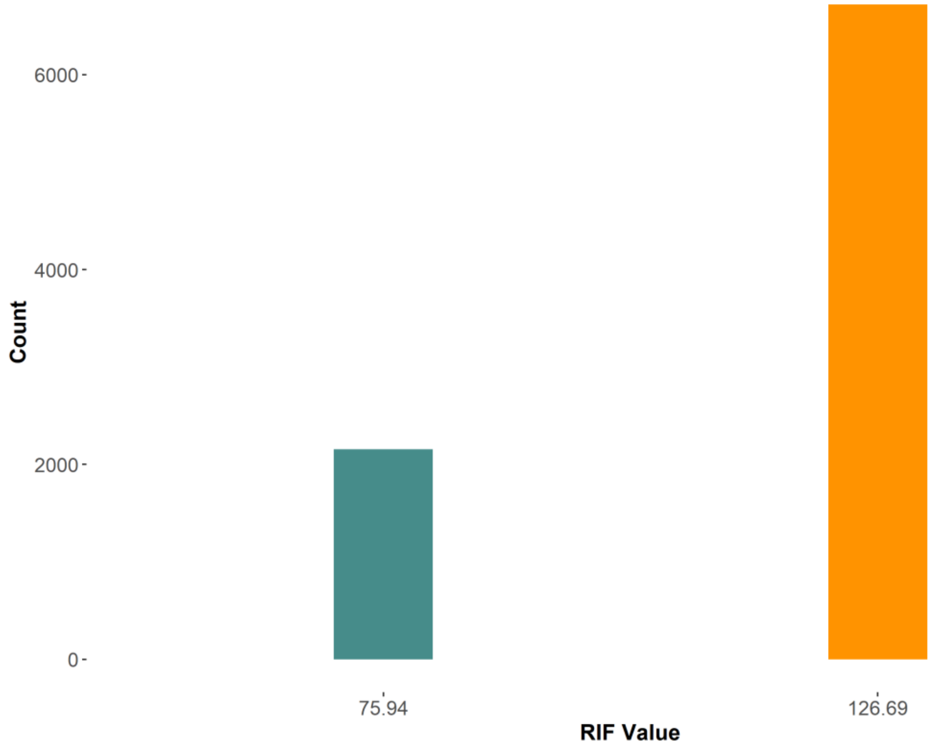
$$E\big[E[RIF(SBP_i; q_{0.25})|X, C]\big] = E\big[\alpha_{0,0.25} + \alpha_{1,0.25}x_i + \lambda'_{0.25}C\big]$$

$$\underbrace{E[RIF(SBP_i; q_{0.25})]}_{E[RIF(SBP_i; q_{0.25})] = q_{0.25}} = \alpha_{0,0.25} + \boxed{\alpha_{1,0.25}}E[X] + \lambda'_{0.25}E[C]$$

The RIF "trick"

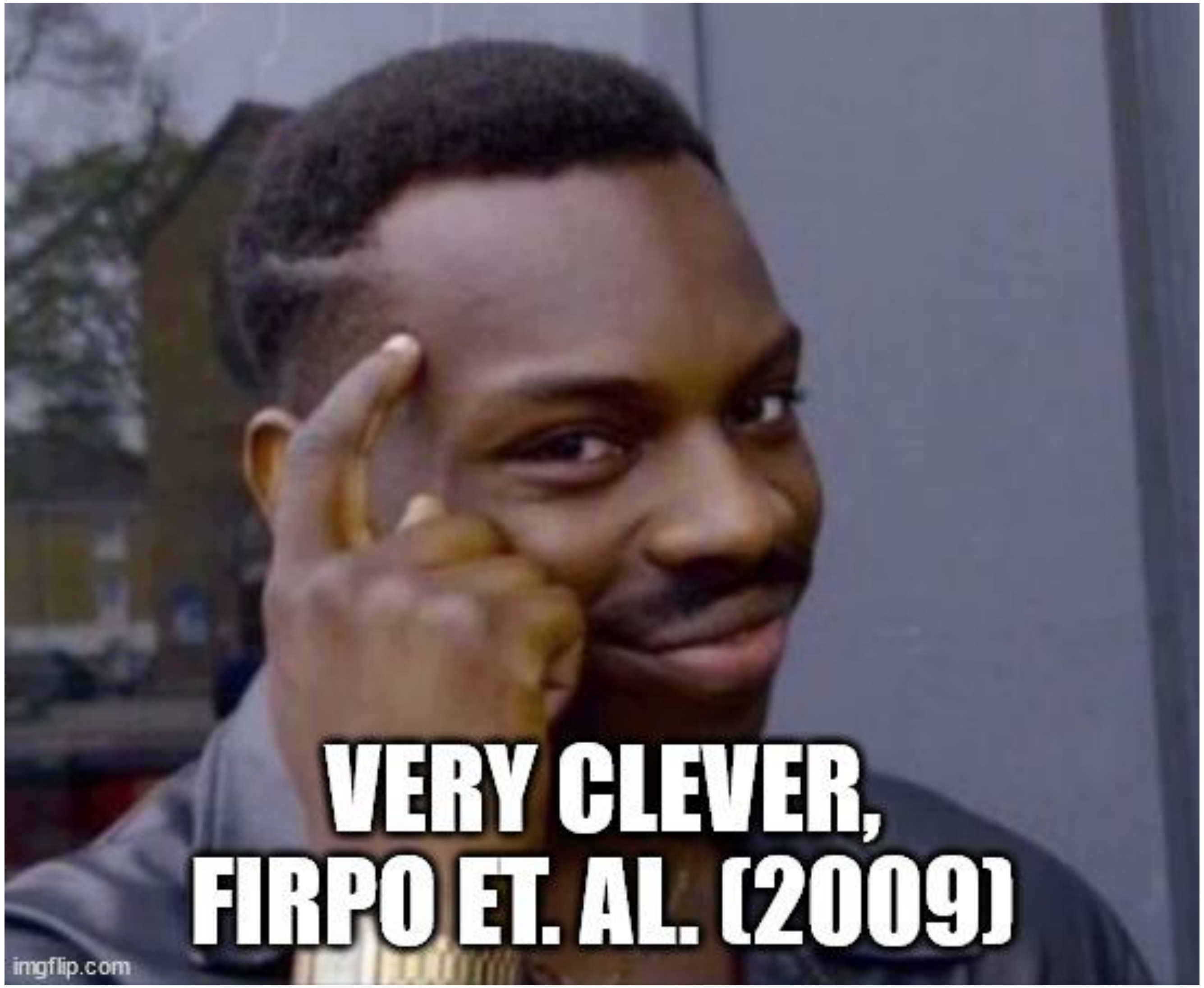
$$RIF(SBP_i, q_{0.25}) = \begin{cases} 75.94; & I(SBP \leq 114) \\ 126.69; & I(SBP > 114) \end{cases}$$





$$E[RIF(SBP_i, q_{0.25})] = (0.25)(75.94) + (0.75)(126.69) = 114\, mmHg = q_{0.25}$$

Firpo’s method “tricks” a regression model into modeling quantiles of the unconditional outcome distribution by using the recentered influence function as the outcome


VERY CLEVER,
FIRPO ET. AL. (2009)
imgflip.com

Interpreting coefficients of the RIF-OLS model

$$E[\widehat{RIF(y_i; q_\tau)}] = \widehat{\alpha_{0,\tau}} + \widehat{\alpha_{1,\tau}} E[X] + \widehat{\lambda_\tau}' E[C]$$

- $\widehat{\alpha_{1,\tau}}$ is the estimated change in the τ^{th} quantile of the unconditional distribution of Y for a unit change in the mean of the unconditional distribution of the exposure X

In the context of our example, unconditional quantile regression answers the question: By how much does the τ^{th} quantile of the unconditional SBP distribution change if the mean of the unconditional education distribution changed by one unit?

Step 1
Estimate the RIF in the unconditional outcome distribution

Step 2
Regress RIF on exposure + covariates

Step 3
Estimate standard errors using bootstrap or sandwich estimator

Step 3: Bootstrap confidence intervals (preferred method)

$$E[RIF(SBP_i;q_{0.25})|X,C] = \alpha_0 + \alpha_1 x_i + \lambda' C$$

$$E\left[E[RIF(SBP_i;q_{0.25})|X,C]\right] = E[\alpha_0 + \alpha_1 x_i + \lambda' C]$$

$$E[RIF(SBP_i;q_{0.25})] = \alpha_0 + \alpha_1 E[X] + \lambda' E[C]$$

Resample 500+ times, fit RIF regression in each sample, then use the 2.5th and 97.5th quantiles of the "sampling distribution"

Step 3: Heteroskedasticity robust standard errors 💪💪💪

$$E[RIF(SBP_i; q_{0.25})|X, C] = \alpha_0 + \alpha_1 x_i + \lambda' C$$

- Estimate heteroskedasticity robust standard errors

$$\widehat{\Omega} = \begin{bmatrix} \widehat{\epsilon_1}^2 & 0 & \cdots & 0 \\ 0 & \widehat{\epsilon_2}^2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & \cdots & \widehat{\epsilon_n}^2 \end{bmatrix} = \begin{bmatrix} (y_1 - \widehat{y_1})^2 & 0 & \cdots & 0 \\ 0 & (y_2 - \widehat{y_2})^2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & \cdots & (y_n - \widehat{y_n})^2 \end{bmatrix}$$

Time for R

Key takeaways

1. Estimating the change in the unconditional quantile using conditional quantile regression estimates is difficult
   - Using alternative methods like Firpo that estimate the unconditional quantile directly is easy

2. The RIF-regression method captures the change in quantiles of the unconditional distribution for a small change in the exposure distribution
   - Since it includes an indicator function, the RIF will only take on two values

3. RIF regression involves regressing the RIF at the quantile of interest on the exposure and other covariates (e.g., using OLS)
   - We are “tricking” a regression to model the unconditional quantiles by using the RIF as the outcome

Vanquishing the *l'homme moyen*

i.e., comparing the different estimators and where do we go from here?

	Linear Regression	Conditional Quantile Regression	Unconditional Quantile Regression
Model	$E[Y\|X] = X'\beta$	$Q_\tau(Y\|X) = X'\beta_\tau,$ $\tau = (0,1)$	$E[RIF(Y; q_\tau)\|X] = X'\beta_\tau,$ $\tau = (0,1)$

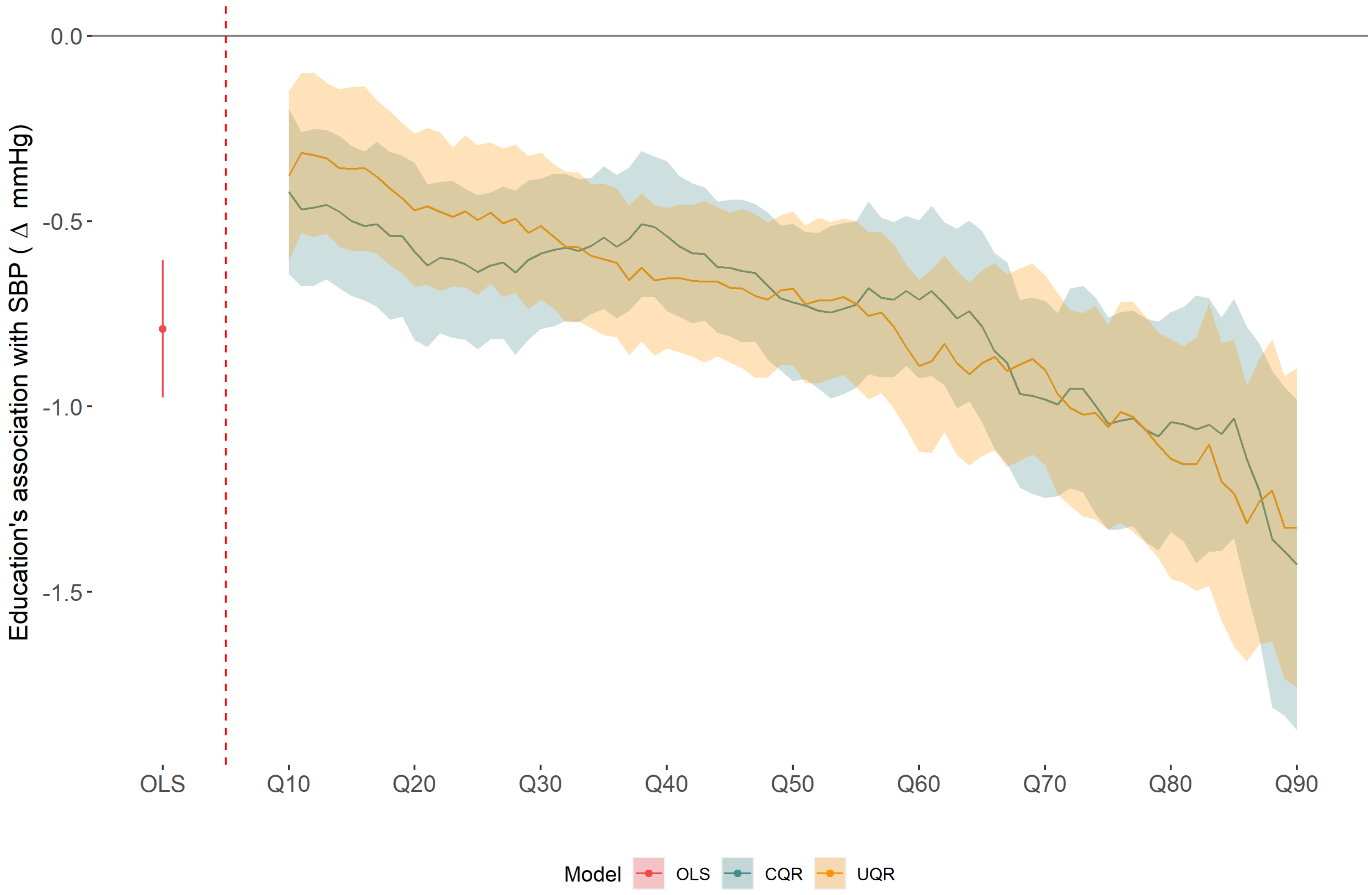
	Linear Regression	Conditional Quantile Regression	Unconditional Quantile Regression
Model	$E[Y\|X] = X'\beta$	$Q_\tau(Y\|X) = X'\beta_\tau$, $\tau = (0,1)$	$E[RIF(Y; q_\tau)\|X] = X'\beta_\tau$, $\tau = (0,1)$
Error	Standard assumptions about the error term (iid, Normal distribution) are usually violated	No distributional assumption is made about the error	No explicit assumption, but likely the same assumptions as the estimation strategy used

	Linear Regression	Conditional Quantile Regression	Unconditional Quantile Regression
Model	$E[Y\|X] = X'\beta$	$Q_\tau(Y\|X) = X'\beta_\tau$, $\tau = (0,1)$	$E[RIF(Y;q_\tau)\|X] = X'\beta_\tau$, $\tau = (0,1)$
Error	Standard assumptions about the error term (iid, Normal distribution) are usually violated	No distributional assumption is made about the error	No explicit assumption, but likely the same assumptions as the estimation strategy used
Estimating coefficients	$\hat{\beta} = (X'X)^{-1}(X'Y)$	$\min_{\hat{\beta}} \sum_{i=1}^{N} \rho_\tau(y_i - x_i'\hat{\beta})$	Method of estimation depends on the estimator used (e.g., RIF-OLS, RIF-Logit)
Solution	Analytic solution	Linear programming methods	

	Linear Regression	Conditional Quantile Regression	Unconditional Quantile Regression
Model	$E[Y\|X] = X'\beta$	$Q_\tau(Y\|X) = X'\beta_\tau$, $\tau = (0,1)$	$E[RIF(Y; q_\tau)\|X] = X'\beta_\tau$, $\tau = (0,1)$
Error	Standard assumptions about the error term (iid, Normal distribution) are usually violated	No distributional assumption is made about the error	No explicit assumption, but likely the same assumptions as the estimation strategy used
Estimating coefficients	$\hat{\beta} = (X'X)^{-1}(X'Y)$	$\min_{\hat{\beta}} \sum_{i=1}^{N} \rho_\tau(y_i - x_i'\hat{\beta})$	Method of estimation depends on the estimator used (e.g., RIF-OLS, RIF-Logit)
Solution	Analytic solution	Linear programming methods	
Estimating standard errors	Several estimators are available + bootstrap	Several estimators are available and key point is that we need to estimate the error density + bootstrap	Same estimators available as would be for the choice of estimator

	Linear Regression	Conditional Quantile Regression	Unconditional Quantile Regression
Model	$E[Y\|X] = X'\beta$	$Q_\tau(Y\|X) = X'\beta_\tau$, $\tau = (0,1)$	$E[RIF(Y;q_\tau)\|X] = X'\beta_\tau$, $\tau = (0,1)$
Error	Standard assumptions about the error term (iid, Normal distribution) are usually violated	No distributional assumption is made about the error	No explicit assumption, but likely the same assumptions as the estimation strategy used
Estimating coefficients	$\hat{\beta} = (X'X)^{-1}(X'Y)$	$\min_{\hat{\beta}} \sum_{i=1}^{N} \rho_\tau(y_i - x_i'\hat{\beta})$	Method of estimation depends on the estimator used (e.g., RIF-OLS, RIF-Logit)
Solution	Analytic solution	Linear programming methods	
Estimating standard errors	Several estimators are available + bootstrap	Several estimators are available and key point is that we need to estimate the error density + bootstrap	Same estimators available as would be for the choice of estimator
Interpretation	One unit increase in the independent variable of interest is associated with a $\hat{\beta}$ unit change in the conditional mean of the outcome	One unit increase in the independent variable of interest is associated with a $\hat{\beta}$ unit change in the τ^{th} quantile of the conditional outcome distribution	One unit increase in the mean of the independent variable of interest is associated with a $\hat{\beta}$ unit change in the τ^{th} quantile of the unconditional outcome distribution

Comparing linear regression, CQR, and UQR estimates



CQR and UQR estimates may be quite different, but even when they are similar, remember that they are estimating quite different estimads

Choosing between CQR and UQR

- What are the aims of your study?
  - Are you interested in making comparisons of the exposure-outcome relationship across groups defined based on individual characteristics? If so, conditional quantiles may be of interest
  - Are you interested in understanding what would happen to the population outcome distribution if everyone had been exposed versus unexposed? If so, unconditional quantiles may be of interest

Choosing between CQR and UQR

- What's the data generating process?
  - Any interactions between the exposure and covariates? If yes, CQR estimates do not generalize to quantiles of the unconditional outcome distribution

- What is your data structure/identification strategy like?
  - More tools developed for using CQR with different data structures, study designs, and data measured with error
  - UQR assumes all covariates included in the model are exogenous (i.e., unconfounded). Newer quantile regression methods targeted at the unconditional quantile may be better suited for endogenous (i.e., confounded) exposures

Extension #1: Conditional quantile regression for longitudinal data

- In longitudinal data, we are concerned with two key sources of variation: within-observation and between-observation

- Standard estimators usually model between-observation variance through random effects models
  - Posits a distribution for the outcome and random effects → use likelihood-based methods to estimate parameters

- CQR makes no assumptions about the shape of the outcome distribution

Extension #1: Conditional quantile regression for longitudinal data

- Geraci and colleagues (2005, 2007, 2014) develop a linear quantile mixed model to estimate CQR in longitudinal data

$$Q_\tau(Y|X,u) = X'\beta_\tau + Z'u$$

Where u is the observation-specific random effects and Z is the design matrix for the random effects

Extension #1: Conditional quantile regression for longitudinal data

- Mixed models are generally solved using likelihood-based methods

- Geraci and colleagues' key insight was to impose the Asymmetric Laplace distribution on the outcome
  - Allowed them to reformulate the linear quantile mixed model estimation problem in likelihood maximization terms
  - Need to choose the appropriate distribution for random effects too

- In R, linear quantile mixed models are implemented using the lqmm package

Extension 2: Causal inference using instrumental variables in CQR

- Standard causal inference assumptions required to estimate "quantile treatment effects"
  - Conditional exchangeability
  - Consistency
  - Positivity

- When the conditional exchangeability assumption is not met, we could use an instrumental variable to estimate causal effect
  - Instrument: a variable which 1) "strongly" affects the exposure; 2) affects the outcome only through the exposure; 3) does not share common causes with the exposure/outcome conditional on covariates

Extension #2: Causal inference using instrumental variables in CQR

- Different estimators are available, all of which make additional assumptions about the variables being modeled:
  1. Abadie et. al. (2002):
     - Identifies the Local Quantile Treatment Effect for a binary instrument/exposure
     - Places additional restrictions on the relationship between the instrument and exposure (“monotonicity assumption”)
  2. Chernozhukov and Hansen (2005):
     - Identifies the Quantile Treatment Effect for any type of instrument/exposure
     - Places additional restrictions on the ranking of individuals across counterfactual distributions of the exposure

Applications of quantile regression at SER 2023

Poster presentations

- Hebert et. al. *Distributional impact of increased post-secondary education on later-life cognition: Evidence from a natural experiment*. Poster session #1, June 13, 7:30-8:30. Poster P47.
- Pederson et. al. *A quantile regression analysis of physical activity and cognition in a racially/ethnically diverse sample of older adults: Results from the Kaiser Healthy Aging.* Poster session #2, June 14, 7:30-8:30. Poster P106.
- Khadka et. al. *Impact of Vietnam-era G.I. Bill eligibility on the distribution of later-life blood pressure: Evidence from a natural experiment.* Poster session #3, June 15, 7:30-8:30. Poster P1305.

Applications of quantile regression at SER 2023

Oral presentations

- Buto et. al. *Heterogeneous associations of HbA1c with MRI measures of brain health.* Epidemiology of neurological impairment across the lifespan. June 14, 3:45-5:15.
- Irish et. al. *Impact of availability of college education on later-life blood pressure distribution: An instrumental variables analysis of a natural experiment.* Novel methods to measure multilevel social factors across the lifecourse. June 15, 10:15-11:45.

Applications of quantile regression at SER 2024: all of you!

Key takeaways

1. Investigating how an exposure affects the entire outcome distribution, in particular the tails, is substantively important

2. Quantile regressions allow us to quantify the relationship of an exposure with the outcome distribution

3. Need to determine if we are interested in quantiles of the conditional or unconditional outcome distribution in advance

4. Separate estimators need to be used for quantiles of the conditional versus unconditional outcome distribution

Thank you!

Aayush Khadka: aayush.khadka@ucsf.edu

Jilly Hebert: jilly.hebert@ucsf.edu

Anusha M. Vable: anusha.vable@ucsf.edu